

White Paper

Grant number: HD-248410-16

Project title: Classical Intertextuality and Computation

Project Director: Pramit Chaudhuri

Grantee institution: The University of Texas at Austin

Submitted: April 30, 2019

PROJECT ACTIVITIES AND ACCOMPLISHMENTS

The Quantitative Criticism Lab (QCL, www.qcrit.org) is a collaborative, cross-disciplinary project to apply methods drawn from the sciences to the study of literature. Co-founded and co-directed by Pramit Chaudhuri (University of Texas at Austin) and Joseph Dexter (Dartmouth College), QCL is partly based at UT Austin but comprises researchers from various institutions, including specialists in philology and literary criticism, biology, and computer science. The major goals for the NEH Start-up grant period were to develop four main computational tools and methods for the analysis of literature, primarily in Latin but also in Ancient Greek and with potential extensions to other languages. The tools and methods provide information about verbal and stylistic relations among texts (“intertextuality”), a core feature of any account of individual literary works or large corpora spanning different periods, regions, and types of production. Such intertextual relations range from close verbal echoes by which one author refers to the work of another to much more general resemblances, such as genres, which define a tradition and structure readers’ responses. The four tools and methods were as follows: 1) a sequence alignment tool, inspired by a core technique in genomics, which identifies verbal parallels that are close but inexact (the commonest kind of intertextuality); 2) a digital Greek-Latin thesaurus to enable identification of parallels across languages by meaning; 3) a set of tools for classification of texts according to various stylistic metrics, especially useful for studies of quotation and attribution; 4) phylogenetic methods to chart the evolutionary histories of classical texts and their traditions of reception. By the end of the grant period three fully functional tools had been created, while one proved to require significant further attention; research demonstrating the application of the methods to major literary questions appeared in journals and conferences across the humanities and sciences, which generated significant media attention; and the winning of major additional grants has ensured the project’s continued expansion, sustainability, and impact.

1) Sequence alignment tool

Our first publicly accessible tool, Fīlum (www.qcrit.org/filum), offers a new method for finding intertexts in Latin literature using sequence alignment, a technique drawn from computational linguistics and bioinformatics. Fīlum is based on sequence alignment of n-grams (combinations of characters or words of arbitrary length) and allows for identification of any short verbal parallel between multiple Latin texts. The method is related to an essential tool for modern biology research, the Basic Local Alignment Search Tool (BLAST), which is used to identify homologous gene or protein sequences. Fīlum allows users to generate lists of n-grams similar to a phrase of interest, ranked according to a distance metric that measures character-by-character similarity (a screenshot in the Appendices shows the interface for Fīlum). Although sequence alignment has been used for textual comparison before, it has been employed at the level of word rather than character. Using character-by-character alignment, however, enables the detection of

phrases that are only partially similar, thereby offering a significant advantage in identifying non-exact parallels, a notoriously tricky yet literarily interesting class of intertextuality. Other prominent word-search and intertextuality detection tools, such as Peter Heslin's Diogenes program, the Italian project Musisque Deoque (www.mqdq.it), and the main tool developed by the Tesserae Project, typically restrict attention to repeated words or phrases and therefore miss such cases.

The project's first major publication, appearing in the classics journal *Dictynna* (2015), uses *Filum* to explore the intertextual connections between Silius Italicus' *Punica* (a Latin epic poem about the Second Punic War fought between Rome and Carthage) and its two main models, Vergil's *Aeneid* and Livy's history of Rome. The article compares the performance of our tool to that of Diogenes and Tesserae and analyzes a series of computationally identified parallels that have not been commented on previously. One strength of sequence alignment is the identification of parallels sharing common sounds, whether morphological endings or more general phonetic resemblances, which helps to address the distinctively aural aspect of poetic composition and fills out the profile of creative variation across texts.

Since the publication of the *Dictynna*, article we have completed a large-scale and systematic test of the effectiveness of the sequence alignment method. Our main source for the current test is Valerius Flaccus' *Argonautica*, an incomplete epic in eight books dating from the 1st century C.E. The *Argonautica* is known to be related to several Greek and Roman predecessors, including Apollonius' *Argonautica* (Greek) and Vergil's *Aeneid* (Latin), which makes Valerius' poem an ideal candidate for testing sequence alignment as an approach to intertextuality detection. The existence of three multiple recent and very full commentaries on Book 1 of the poem provides a solid basis for assessment of digital tools' capacities for generating new intertexts and capturing known ones. In order to demonstrate the efficiency of using sequence alignment to identify intertexts in a substantial corpus, we selected a group of fifteen Latin poems plausibly connected with Valerius' *Argonautica* (five long epics, including Valerius' poem itself, and ten tragedies) and assembled a database of 1,300 parallels noted in the commentaries. Of the intertexts catalogued, 84% were recovered using *Filum* with 100 or fewer off-target results, along with over 250 previously unnoticed intertexts of potential literary significance. With a view to demonstrating the potential use of sequence alignment for *de novo* analysis, especially in subfields that enjoy relatively little scholarly attention, we have also applied the tool to Maffeo Vegio's supplement to Vergil's *Aeneid*, a neo-Latin work that completes the story of Aeneas' war in Italy. The validation data for Book 1 of the *Argonautica* and Maffeo Vegio's *Supplementum* represents the major part of an article designed for an interdisciplinary science journal, which will be submitted in 2019.

We have also developed a modified version of the sequence alignment tool that allows for specific detection of anagrams. Though relatively rare across the Latin corpus, anagrammatic

wordplay has particular importance for certain authors (e.g., Lucretius) and is deployed more infrequently but in literarily significant ways by others (e.g., Vergil). Such deliberately crafted anagrams pose a double challenge to the critic: first in identifying them, and then in establishing the intentionality of the wordplay. Since few if any critics are likely to search for anagrams systematically using “manual” reading alone, detection typically relies upon chance observations made in the course of studying a particular text. Computation makes the former task vastly more efficient and enables rapid analysis of entire works and even systematic study of the Latin corpus. The use of an approach rooted in sequence alignment enables detection of both exact and inexact anagrams. In the course of testing the prototype anagram search tool, the Co-PIs identified two new and significant anagrams in a sample passage from *Aeneid* 8, which advances prior discussions of other well-known anagrams from the same passage. The two examples include an exact and inexact anagram, both of which embody a larger thematic in the passage concerned with the composition, decomposition, and recomposition of societies over time. The case study thus highlights the affinity between methodological development and literary interpretation that motivates the project as a whole. Other discoveries include identification of new literarily significant anagrams in Lucretius, an author known for exploiting this and related forms of wordplay elsewhere in his text. Although Latin word play has received substantial and recent scholarly attention, the tool offers significantly enhanced capabilities to critics working in this area and drastically reduces the barrier to entry for any critic interested in casual exploration of such technical phenomena. We plan to make the tool available via the QCL website coincident with publication of the paper presenting the associated data.

2) Greek-Latin thesaurus

Latin authors of all genres and periods were deeply attentive to their Greek antecedents, and certain classes of interlinguistic intertextuality, such as between Hellenistic and Augustan poetry, have been among the most intensively studied using traditional critical methods. No general computational methods, however, are currently available for the analysis of Greek-Latin intertextuality. As part of the Start-up grant, we sought to develop a method for cross-linguistic intertextual search that exploits early modern Greek-to-Latin translations of canonical texts to create a “digital thesaurus” for identifying thematically similar but lexically distinct phrases. Working with a large team of research assistants, we successfully assembled a custom thesaurus containing over 8,500 unique Latin words with one or more Greek equivalents, drawn from bilingual editions of Homer’s *Iliad* and *Odyssey* and the 11 extant comedies of Aristophanes, as well as from the *Colloquia of the Hermeneumata Pseudodositheana*.

We then investigated the usefulness of our thesaurus for general-purpose Greek/Latin intertextual search. In preliminary research, we demonstrated that integrating the thesaurus with the Tesserae approach for identifying and scoring intertextual parallels enabled us to identify

obvious cross-linguistic parallels (such as very common Greek phrases translated by Roman authors). Despite significant optimization, however, our tool struggled to find intertexts of greater literary interest, perhaps due to sparseness of the thesaurus or non-generalizability of the scoring metric from single-language to cross-linguistic search. Given these technical challenges and the opportunity for unexpected expansion of our research agenda on stylometry (see below), we opted to prioritize other aims of the Start-up grant ahead of developing a public Greek-Latin search tool.

3) Stylometry

To date we have created a set of tools to tackle a specific subset of problems in understanding literary style and text reuse. These tools calculate the frequency of certain features in a text (e.g., character n-grams, non-content words, relative clauses, and, for poetic texts, enjambments). Variation in the presence of these and other features can reflect stylistic differences either within a single text or across a range of texts and thereby help to identify anomalous passages. A core goal of our work is to adapt such stylometric techniques to address literary critical concerns that are subtler than binary attribution questions. In our research we have applied the tools to two related problems in classical literature: stylometric profiling 1) of corpora containing imitative works with a view to identifying distinguishing features of literary interest, and 2) of works containing large passages of quotation or paraphrase.

For the former, we examined a range of features across tragedies attributed to Seneca or composed by authors imitating Seneca. Noteworthy results include the remarkable propensity for enjambment (much higher than the Senecan average) observable in the *Progne*, a neo-Latin play composed by Gregorio Correr partly in imitation of Seneca's *Thyestes*. Although impossible to explain with certainty, we venture that the statistical anomaly reflects the desire of the young but highly talented Correr - only 18 at the time - to exhibit a flexibility in the handling of the verse line characteristic of a more mature poet. Another striking result pertains to Seneca's incomplete *Phoenissae*, a tragedy on the civil war between the twin sons of Oedipus. N-gram analyses show a disproportionate clustering of certain morphological endings in nearby lines. Building on prior claims made for the overlap between form and content in the play, we suggest that the sound clusters reflect the *Phoenissae*'s pervasive interest in themes of twinning and repetition.

In ancient literature, and especially in historiography, it is often unclear whether a particular passage should be thought of as direct quotation or paraphrase, and, if paraphrase, to what extent the passage exhibits the stylistic features of the source. As a result, modern commentators often differ widely in their assessments of the authorship of passages of "quotation." We have used the microscopic analyses of style enabled by computation to provide a more robust undergirding for philological arguments concerning such disputed passages. In particular, we sought to discover whether the body of citations in Livy's Roman history can be computationally discriminated in

order to generate a “thermal map” of their Livian or non-Livian qualities. Use of a one-class support vector machine (SVM) with a broad set of stylometric features achieved a remarkably clear distinction between the passages of quotation/paraphrase and the remainder of Livy. We then compared Livy with a wide selection of prose and poetic texts across Roman literary history. The data show an expected affinity among historiographical prose texts and among post-Republican prose texts, but a marked difference from (unsurprisingly) poetry and (more interestingly) Ciceronian prose. This research on Senecan and Livian style was published in 2017 in the *Proceedings of the National Academy of Sciences*, a high-profile interdisciplinary science journal.

An example application of the toolkit is the set of experiments reported in our paper recently published in *Digital Scholarship in the Humanities*. We sought to measure a large number of stylometric features in Latin prose and verse and then use those data to train machine learning classifiers that could distinguish the two classes. Although the distinction is a relatively coarse one for any human reader to discern, our intention was to build an effective feature set that not only would achieve the immediate goals of the experiment but could also be developed and refined for future studies of corpora that are larger and more heterogeneous. With 26 features calculated for over 700 Latin literary texts using the toolkit, the classifiers performed extremely well (with five-fold cross-validation, accuracy values of 97.6 ± 1.1 % for a random forest and 97.8 ± 1.4 % for an SVM with a linear kernel). Moreover, we performed statistical feature ranking to identify the stylistic characteristics most useful for differentiating prose and verse. While no single feature is crucial to the success of the classifier, our systematic ranking identifies features associated with hypotaxis (such as sentences containing relative clauses) and with metrical constraints (superlatives) to be important.

To explore the use of stylometry in other premodern traditions, in early 2017 we entered into a collaboration with several specialists in Old English (OE) literature. Together we constructed the first corpus-wide stylometric profile of OE verse, incorporating both features drawn from our work on Latin poetry (sense-pauses and character n-grams) and language-specific features such as nominal compounds. This profile, which provides an intricate diachronic portrait of the earliest development of English as a literary language, enabled us to address seminal problems in OE philology (e.g., regarding the stylistic continuity of *Beowulf*) and should be of value for future quantitative analyses of the cultural evolution of English literature. A collaborative paper describing the analysis has recently been published in the journal *Nature Human Behaviour*; among the most interesting philological results reported therein is a striking similarity in the use of nominal compounds between the unsigned poem *Andreas* and the works of Cynewulf. This surprising result suggests either that Cynewulf (the first author to whom multiple English poems can be ascribed) wrote *Andreas*, or that its creator was influenced profoundly by a Cynewulfian school of poetry.

4) Phylogenetic profiling

We have developed an approach, termed phylogenetic profiling, for visualizing large-scale relationships among versions of the same story or works within a tightly-knit tradition. These relationships are typically categorized and studied by classical scholars under the rubric of “classical reception,” a term denoting the adaptation and re-imagining of Greco-Roman works in new cultural contexts. As such, reception typically involves transmission and transformation over long periods and therefore invites a particularly evolutionary perspective. Phylogenetic profiling is a standard approach in computational biology to identify proteins that are functionally associated across long evolutionary histories. Building on recent interest in the interplay between bioinformatics and literary study, we have begun to develop phylogenetic profiling as a tool for humanistic study.

Our initial application of phylogenetic profiling has been to the reception of classical drama. In our approach, each text in the reception tradition is treated as an organism, and we determine the binary presence or absence of characters in each text. As with biological phylogenetic profiling, our approach is useful for elucidating thematic relationships undetectable through individual intertextual searches and provides a framework for visualizing and interpreting complex reception histories. As part of the grant work, phylogenetic profiles based on character lists have been produced for Aeschylus’ *Agamemnon*, Sophocles’ *Antigone*, Euripides’ *Heracles*, Plautus’ *Amphitryon* and their respective reception traditions (which can include in excess of 100 adaptations).

AUDIENCES

The Co-PIs have undertaken a wide variety of formal and informal activities to disseminate the results of the grant work to diverse audiences. Formal activities include five invited lectures to scholarly audiences at institutions including Yale University, the University of Michigan, and the University of Iowa, as well as four presentations at the annual meeting of the Society for Classical Studies (the largest and most important conference within Classics). Of particular interest, two of these presentations were interactive, hands-on demonstrations of QCL tools conducted as part of the Ancient MakerSpaces workshop, which was initiated in 2017. In addition, extensive participation of students of all career stages in project research has ensured broad dissemination to the next generation of researchers. Three graduate students at UT Austin participated in work related to the Start-up grant, along with numerous undergraduate students from UT Austin, Dartmouth, and other institutions and high school students recruited through the Research Science Institute. Four of the six publications produced to date have at least one undergraduate or high school student co-author (see Appendix for full list).

Beyond dissemination to the immediate scholarly community, grant research has generated extensive media coverage, including popular articles in the *Guardian*, *Times*, *Boston Globe*, and *Harvard Medicine Magazine* (see below for details). In addition, the Co-PI has developed a new course for Dartmouth undergraduates on quantitative literary criticism, which incorporates Filum and the stylometry toolkit as well as many of the grant publications. He is teaching the course during the Spring 2019 term.

EVALUATION

At present, insufficient time has elapsed since the completion of the project to fully evaluate the project tools or their uptake in the wider scholarly community. We continue to explore and implement enhancements to the work performed as part of the Start-up grant. The lengthy publication cycle, especially in the humanities, has meant that citations of our work in humanities venues are limited, though some are already in evidence. In the sciences, where the cycle is shorter, the paper in *PNAS* has already been cited at least seven times. The most revealing measure of success available at the moment is publication itself and the response in the wider media. The aim to publish the grant-funded research in diverse venues spanning the humanities and sciences has been accomplished; in two cases, moreover, the venues have been very prestigious: *PNAS*, as mentioned, and *Nature Human Behaviour* for the spin-off work on Old English. These high profile, cross-disciplinary journals for research in science and social science rarely publish work in the humanities; placement of these two articles thus resulted in dissemination of digital humanities research to an unusually wide audience. *PNAS*, in particular, drew attention to our work by selecting it for inclusion in the “In this issue” section of the print volume.

The project has gained recognition in the national and international media. It was one of two NEH projects featured by the Canadian Broadcasting Corporation as part of a story on threats to federal funding for the arts, humanities, and cultural programs. The *PNAS* paper led to features in two university publications, UT News and Harvard Medicine Magazine. The article in *Nature Human Behaviour* generated an especially large response in the international media, including stories in all the major British newspapers, an editorial in the print edition of the *Times*, as well as newspapers, magazines, and internet publications in the United States and Germany.

CONTINUATION

The project’s numerous successes - especially in terms of published research appearing in high profile venues across diverse disciplines, media attention, and fund-raising - have encouraged its core participants to commit greater time towards its goals and have attracted a wider circle of collaborators. The Co-PI, Dr. Joseph Dexter, has since completed his Ph.D in Systems Biology

and has obtained two prestigious independent postdoctoral fellowships in data science, at Dartmouth (2018-19) and Harvard University (2020-21). These positions will enable him to devote a significant proportion of his research portfolio to the work, while the PD, Dr. Pramit Chaudhuri, continues to make this research his top priority at UT Austin. The winning of a Digital Extension Grant from the American Council of Learned Societies in 2018 enabled the recruitment of a postdoctoral fellow, Dr. Patrick Burns, who has significant experience in classical literary scholarship and the creation of text processing and analytical tools for under-resourced languages. The project team has applied for an NEH Digital Advancement Grant to extend their work into non-Western languages and to implement a computational protocol for the phylogenetic profiling methods developed as part of the Start-Up grant work. The research to date has also created and strengthened collaborations that will be a key part of future work, especially in Old English (Dr. Madison Krieger at Harvard University) and Bengali (Dr. Sukanta Chaudhuri at Jadavpur University). Finally, the PD's previous position at Dartmouth and the Co-PI's current fellowship there, along with the continued collaboration with Dartmouth's Research Computing unit, have built a substantial connection between researchers at UT Austin and Dartmouth. This connection is most fully realized in the two-part conference described in the section below.

LONG-TERM IMPACT

The main spin-off program resulting from the project is the expansion of research into Old English and other traditions beyond Latin and Greek. Although such extensibility had always been a long-term aim of the project, the rapidity and scale of success, as demonstrated by the publication in *Nature Human Behaviour* and the resulting attention from the media, including almost every high profile newspaper in the UK, could not have been anticipated.

Cross-pollination of this research across languages also drove the two-part conference "Digital Humanities Beyond Modern English: Computational Analysis of Pre-modern and Non-Western Literature," the first part of which was held in April 2019 at Dartmouth and attracted major support from the Neukom Institute for Computational Science and the Leslie Center for the Humanities. The event brought together over 15 core participants who work on a diverse array of premodern and non-Anglophone traditions, including Latin, Ancient Greek, Coptic, Old English, Celtic, Sanskrit, Bengali, Spanish, and Chinese. The Dartmouth event paves the way for the second part of the conference to be held at UT Austin in 2020, which is partly funded by a Digital Extension Grant from the American Council of Learned Societies awarded in 2018. That same grant also supports significant enhancements of Fīlum and the stylometry toolkit. The conference is intended to result in an edited volume that will seed the types of research developed during the Start-up grant among a diverse group of literary and digital humanities scholars working across languages and periods.

AWARD PRODUCTS

The major outputs from the grant comprise computational tools, published scholarship, and research presentations, all of which are listed below. All tools and code relating to published research are available from the QCL website and Github repository. As further research is published, associated tools will be made available on the project website and code will be released via the repo. Where possible QCL has paid for published papers to be open access, and papers are listed on the project website with links to supporting information where relevant. All papers relating to the project research cite the grant number and the support of the NEH in the acknowledgements.

Project website and code repository:

1. Project information and tools at: www.qcrit.org.
2. Repository for code relating to published papers: www.github.com/qcrit

List of publications:

1. J.P. Dexter, T. Katz, N. Tripuraneni, T. Dasgupta, A. Kannan, J.A. Brofos, J.A. Bonilla Lopez, L. Schroeder, A. Casarez, M. Rabinovich, A. Haimson Lushkov, and P. Chaudhuri, "Quantitative criticism of literary relationships," *Proceedings of the National Academy of Sciences USA* **114** (2017) E3195-E3204
2. P. Chaudhuri and J.P. Dexter, "Bioinformatics and Classical Literary Study," *Journal of Data Mining and Digital Humanities* (2017) <http://jdmdh.episciences.org/3807>
3. J.P. Dexter, K. Iyer, T. Dasgupta, and P. Chaudhuri, "A small set of stylometric features differentiates Latin prose and verse," *Digital Scholarship in the Humanities* (2018) **doi:10.1093/llc/fqy070**
4. L. Neidorf, M. S. Krieger, M. Yakubek, P. Chaudhuri, and J.P. Dexter, "Large-scale quantitative profiling of the Old English verse tradition," *Nature Human Behaviour* (2019) **doi:10.1038/s41562-019-0570-1**
5. T. Gianitsos, T.J. Bolt, P. Chaudhuri, and J.P. Dexter, "Stylometric Classification of Ancient Greek Literary Texts by Genre," *LaTeCH-CLfL 2019: The 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature* (Forthcoming 2019)
6. J.P. Dexter and P. Chaudhuri, "*Dardanio Anchisae*: Hiatus, Homer, and Intermetricality in the *Aeneid*," *Harvard Studies in Classical Philology* **111** (Forthcoming 2019)

List of presentations:

1. P. Chaudhuri and J.P. Dexter, “Quantitative Criticism of Classical Literature,” Department of Classics, University of Iowa, November 2016
2. J.P. Dexter, P. Chaudhuri, and A. Schwartz, “Phylogenetic Profiling and the Reception of Classical Drama,” Ancient MakerSpaces: Digital Tools for Classical Scholarship, 148th Annual Meeting of the Society for Classical Studies, Toronto, January 2017
3. T.J. Bolt, A. Casarez, and J.H. Flynt, “How to Do Philology with Computers,” Ancient MakerSpaces: Digital Tools for Classical Scholarship, 149th Annual Meeting of the Society for Classical Studies, Boston, January 2018
4. P. Chaudhuri and J.P. Dexter, “More Latian Anagrams (*Aen.* 8.314-36),” 149th Annual Meeting of the Society for Classical Studies, Boston, January 2018
5. P. Chaudhuri, “Family Resemblances: Computational Profiling of Silver Latin and its Early Modern reception,” Graduate Elected Speaker, Department of Classics, Yale University, April 2018
6. P. Chaudhuri and J.P. Dexter, “The Ship of Theseus: A framework for intertextuality connecting literature, biology, and computation,” Digital Classics Association Panel on “Reconnecting the Classics,” 150th Annual Meeting of the Society for Classical Studies, San Diego, January 2019
7. J.P. Dexter, “Quantifying Literary Style and Evolution,” Department of Computer Science and Statistics, University of Rhode Island, February 2019
8. J.P. Dexter, “Quantifying Literary Style and Evolution,” Department of Statistics and Data Science, Yale University, March 2019
9. P. Chaudhuri, “Digital methods for Latin literary study: a Quantitative Criticism Lab workshop,” University of Michigan, March 2019

APPENDICES

Screenshots:

- Filum interface
- Stylometry toolkit interface

Representative papers:

- J.P. Dexter, T. Katz, N. Tripuraneni, T. Dasgupta, A. Kannan, J.A. Brofos, J.A. Bonilla Lopez, L. Schroeder, A. Casarez, M. Rabinovich, A. Haimson Lushkov, and P. Chaudhuri, “Quantitative criticism of literary relationships,” *Proceedings of the National Academy of Sciences USA* 114 (2017) E3195-E3204
- J.P. Dexter, K. Iyer, T. Dasgupta, and P. Chaudhuri, “A small set of stylometric features differentiates Latin prose and verse,” *Digital Scholarship in the Humanities* (2018) doi:10.1093/llc/fqy070
- L. Neidorf, M. S. Krieger, M. Yakubek, P. Chaudhuri, and J.P. Dexter, “Large-scale quantitative profiling of the Old English verse tradition,” *Nature Human Behaviour* (2019) doi:10.1038/s41562-019-0570-1

Samples of media coverage:

- “Across America, artists are searching for answers about Trump’s planned funding cuts,” by Haydn Watters, *CBC News*, March 26, 2017
- “A Closer Read,” by Kevin Jiang, *Harvard Medicine Magazine*, November 17, 2017
- “Beowulf the work of single author, research suggests,” by Nicola Davis, *The Guardian*, April 8, 2019
- “‘Beowulf’ is bloody, canonical, and long — and one person wrote it, scholars say,” by Travis Andersen, *The Boston Globe*, April 11, 2019

QUERY

ENTER SEARCH PHRASE

*Enter words in the QUERY box. The query should typically be 2-4 words, each separated by a single space.

MAXIMUM DIFFERENCE

MAX # OF INTERVENING WORDS
Click Here To Enable Order-Free Search

TEXT SELECTION

AUTHOR	TEXT	BOOK
--------	------	------

Reset

Submit

Screenshot of Filum interface

TEXT SELECTION

Select texts to analyze using the drop-down menus. You may select an author (to analyze all associated works), a whole text, or a specific book. Once you have made a selection, click Add. Multiple selections are allowed. Use the Advanced Options menu to save custom corpora and to upload additional texts.

AUTHOR

TEXT

BOOK

☐ SELECT ALL TEXTS

Advanced Options

+ Add

SELECTED TEXTS

This table shows the works selected from the Text Selection menu. Click the Remove button to delete a selection.

Author	Text	Book	Action
--------	------	------	--------

FEATURE SELECTION

Select the stylometric features to calculate. Multiple selections are allowed (within and across categories).

Pronouns and Non-Content Adjectives

SELECT

Subordinate Clauses

SELECT

Conjunctions

SELECT

Miscellaneous

SELECT

☐ Select All Features

Reset

Submit

Screenshot of stylometry toolkit interface

Quantitative criticism of literary relationships

Joseph P. Dexter^{a,1,2}, Theodore Katz^{b,c,d,1}, Nilesh Tripuraneni^{e,1}, Tathagata Dasgupta^{a,1}, Ajay Kannan^f, James A. Brofos^f, Jorge A. Bonilla Lopez^g, Lea A. Schroeder^g, Adriana Casarez^h, Maxim Rabinovichⁱ, Ayelet Haimson Lushkov^j, and Pramit Chaudhuri^{g,i,2}

^aDepartment of Systems Biology, Harvard Medical School, Boston, MA 02115; ^bThe Dalton School, New York, NY 10128; ^cResearch Science Institute, Center for Excellence in Education, McClean, VA 22102; ^dDepartment of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139; ^eDepartment of Engineering, University of Cambridge, Cambridge CB2 1PZ, United Kingdom; ^fDepartment of Computer Science, Dartmouth College, Hanover, NH 03755; ^gDepartment of Classics, Dartmouth College, Hanover, NH 03755; ^hAustin Independent School District, Austin, TX 78703; ⁱDepartment of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720; and ^jDepartment of Classics, University of Texas, Austin, TX 78712

Edited by Kenneth W. Wachter, University of California, Berkeley, CA, and approved February 27, 2017 (received for review July 20, 2016)

Authors often convey meaning by referring to or imitating prior works of literature, a process that creates complex networks of literary relationships ("intertextuality") and contributes to cultural evolution. In this paper, we use techniques from stylometry and machine learning to address subjective literary critical questions about Latin literature, a corpus marked by an extraordinary concentration of intertextuality. Our work, which we term "quantitative criticism," focuses on case studies involving two influential Roman authors, the playwright Seneca and the historian Livy. We find that four plays related to but distinct from Seneca's main writings are differentiated from the rest of the corpus by subtle but important stylistic features. We offer literary interpretations of the significance of these anomalies, providing quantitative data in support of hypotheses about the use of unusual formal features and the interplay between sound and meaning. The second part of the paper describes a machine-learning approach to the identification and analysis of citational material that Livy loosely appropriated from earlier sources. We extend our approach to map the stylistic topography of Latin prose, identifying the writings of Caesar and his near-contemporary Livy as an inflection point in the development of Latin prose style. In total, our results reflect the integration of computational and humanistic methods to investigate a diverse range of literary questions.

authorship attribution | cultural evolution | intertextuality | machine learning | stylometry

The study of literature relies on mapping interactions between texts. Ancient Greek critics understood the tragedies of Aeschylus in part through their relation to Homeric epic, and ancient Roman commentators interpreted words and phrases in texts by citing parallels in other works. Much of literary criticism today rests on understanding these vast networks of intertextuality, which often have profound consequences for the meaning of both individual texts and larger groupings by genre or period (1). Through quantitative analysis of formal elements and their change over time, the study of intertextuality can shed light on the cultural evolution of literature (2).

A central challenge in the study of intertextuality is its heterogeneous nature. Literary parallels differ widely in both similarity and scope (Fig. 1A). The relationship between the associated texts can range from obvious (direct quotation) to extremely subtle (artfully constructed indirect references, often referred to as allusions in literary study). Furthermore, parallels can operate on the level of individual words or phrases, short passages, or entire works and can involve verbal, syntactic, phonetic, or metrical features. As illustrated in Fig. 1A, intertexts can be of comparable similarity but very different scope; an adaptation of an entire work, for instance, can be thought of as a collection of many (local) allusions.

In this paper, we focus on the quantitative characterization of intertextual relationships that involve some (but not exten-

sive) similarity between the works. We take as a case study two problems in classical Latin literature that are of substantial current interest to literary critics and historians. The literature of the Roman Republic and Empire contains an extraordinary density and diversity of intertextual parallels. Intertextuality has become an essential focus of modern critics of Latin literature, and detailed qualitative taxonomies of Latin intertextuality have been constructed (3–5). Another advantage of our focus on classical literature is the near-complete digitization of extant texts in searchable, high-quality databases (6).

It has been a longstanding goal of research in the digital humanities to integrate quantitative methods with the aims of literary study. Following the lead of Burrows' 1987 book *Computation into Criticism*, more recent attempts have involved the theorization and implementation of methods of "distant reading" (7, 8), "algorithmic criticism" (9), "macroanalysis" (10), and "literary pattern recognition" (11). This work has been augmented by additional theoretical analyses (12, 13) and empirical studies that exploit specific methodological innovations, such as topic modeling, often for large-scale profiling of genres or periods (10, 14, 15). Quantitative methods have been especially valuable for the characterization of intertextuality both classical and modern. Computational searches for lexically similar phrases,

Significance

Famous works of literature can serve as cultural touchstones, inviting creative adaptations in subsequent writing. To understand a poem, play, or novel, critics often catalog and analyze these intertextual relationships. The study of such relationships is challenging because intertextuality can take many forms, from direct quotation to literary imitation. Here, we show that techniques from authorship attribution studies, including stylometry and machine learning, can shed light on inexact literary relationships involving little explicit text reuse. We trace the evolution of features not tied to individual words across diverse corpora and provide statistical evidence to support interpretive hypotheses of literary critical interest. The significance of this approach is the integration of quantitative and humanistic methods to address aspects of cultural evolution.

Author contributions: J.P.D., T.K., N.T., T.D., M.R., A.H.L., and P.C. designed research; J.P.D., T.K., N.T., A.K., J.A.B., J.A.B.L., L.A.S., A.C., and P.C. performed research; J.P.D., T.K., N.T., T.D., A.K., J.A.B., J.A.B.L., L.A.S., A.C., A.H.L., and P.C. analyzed data; and J.P.D., A.H.L., and P.C. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

¹J.P.D., T.K., N.T., and T.D. contributed equally to this work.

²To whom correspondence may be addressed. Email: jdexter@fas.harvard.edu or pramit.chaudhuri@austin.utexas.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1611910114/-DCSupplemental.

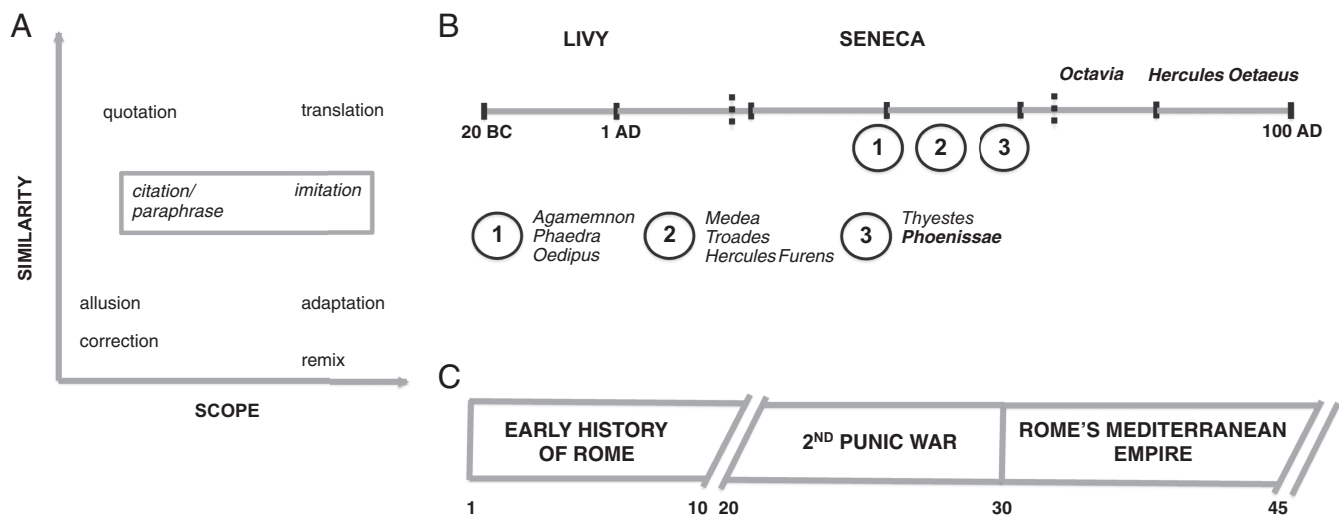


Fig. 1. Intertextuality in Seneca and Livy. (A) Categories of intertextuality. Instances of intertextuality can be characterized according to the similarity between the source text and intertext and the scope of the association. For instance, a short quotation (upper left) exhibits higher similarity and narrower scope than a loose adaptation of an entire play (lower right). The primary focus of the paper is imitation of Seneca and citation/paraphrase in Livy (gray box). (B) Timeline indicating the dates of composition of the texts analyzed. The eight tragedies of Seneca are often divided into early (1), middle (2), and late (3) groups. The two pseudo-Senecan tragedies were composed shortly after his death. Dotted lines indicate the dates of death of Livy and Seneca. (C) Schematic of Livy's history of Rome, which contained 142 books. Books 11–20 and 46–142 have been lost; the subject matter of the surviving books is summarized.

exemplified by the work of the Tesseract and Perseus projects on Greek and Latin literature, are useful for the high-throughput identification of local verbal intertexts (16–19). Such work was highlighted in a 2016 special issue of the journal *Digital Humanities Quarterly* devoted entirely to digital methods and classical studies (20). Digitization of enormous corpora, such as Google Books and the Project Gutenberg Digital Library, has enabled “culturomic” analyses of global linguistic trends (21–24). A notable recent application of such methods was a large-scale study of stylistic influence in English literature based on use patterns of “content-free” words (25). Finally, quantitative stylistic analyses have long been used to clarify gross relationships between texts. Standard applications of stylometry include dating literary works and resolving questions of attribution (26–30). Both ad hoc stylometric analysis and supervised machine learning with stylometric features have proven successful for such applications (31–33), including for cases in Latin literature (34).

Whether an entire work is spurious or authentic, however, is a coarser question than typically posed in literary criticism. Of greater interest is how the spurious work differs from authentic writings and how its composition was influenced by the larger tradition. Recent studies have begun to repurpose stylometry to answer such literary critical questions (10, 35–39). Much of this research relies on the suitability of techniques of authorship attribution for addressing broader literary questions (40). Here, we show that complex relationships between partially similar texts, exemplified at short scales by literary paraphrase and large scales by creative imitation of entire works, can be characterized through the application of stylometry and machine learning, core methods in computational attribution studies. Although the authorship of most of the texts under consideration is not in dispute, these methods allow us to characterize similarities and differences between them in great detail. Our experiments thus provide a richer profile of known intertextual relationships by showing continuity of certain stylometric features within a tradition as well as individual or collective departures from that tradition, and by enabling exploration of the interplay between style and theme.

Although much work in computational text analysis has focused on the word or phrase as the principal unit of analysis, some recent research has shown the utility of other kinds of units, such as character and rhythm, in both large- and small-scale quantitative analyses of literature (41, 42). Our work quantifies a selection of subverbal, syntactic, and prosodic features, which have also been used for authorship attribution. We redeploy these techniques to resolve multiple literary problems of interest to classicists and other humanists.

The philosopher and statesman Seneca (4 BC to AD 65) (Fig. 1B) wrote tragic plays, 10 of which have been transmitted under his name via the medieval manuscript tradition and hugely influenced later dramatists, such as Shakespeare and Racine (43, 44); 2 of these 10 (the *Octavia* and the *Hercules Oetaeus*) are spurious, however, the work of careful imitators writing in the years after Seneca's death. Despite considerable attention, the precise literary and stylistic relationships among both the 8 works attributed to Seneca and the entire corpus of 10 transmitted texts remain unclear. Our computational analysis identifies several subtle but significant differences in poetic style between the *Octavia* and the *Hercules Oetaeus* and the eight authentic tragedies. We extend these methods to contrast typical Senecan style with that of the *Procne*, a neo-Latin tragedy influenced by Seneca but written centuries after his death, and the *Phoenissae*, an authentic but incomplete play. Although easily tabulated computationally, the differentiating features cannot be studied using traditional means without substantial repetitive effort.

The historian Livy (64 or 59 BC to AD 17) (Fig. 1C) wrote a monumental history of Rome covering the period from the city's foundation and the rise of the Roman empire to Livy's contemporary world. The work consisted of 142 books (~2 million words), of which only 35 survive. Livy makes frequent reference to previous works of history, but his citational practices are poorly understood. He cites and quotes both named and unnamed sources, he blends paraphrase and direct quotation, and he freely composes passages in ways likely informed by his reading of sources (45). This complex combination of text reuse has posed particular challenges for literary critics seeking to understand Livy's relationship to his sources. We use an anomaly

detection algorithm trained with a set of 25 stylistic features to classify most material in a curated database of possible citations as differing in style from the rest of Livy. We then apply a similar method to profile the development of Latin prose style across several centuries, which identifies the histories of Caesar and Livy as marking the start of a pronounced shift in literary style that extends across multiple genres.

Results

Quantitative Criticism Identifies Literary Differences Across the Senecan Corpus and Tradition. We profiled a broad range of stylistic features across the whole Senecan and pseudo-Senecan corpus and in Gregorio Correr's *Procne*, a 15th century neo-Latin tragedy deeply influenced by Seneca. We first considered sense pauses (interruptions in speech indicated by any punctuation mark other than a comma), which have proven useful in manual studies of Senecan style. We observed almost no variation in the length-normalized number of sense pauses across the eight authentic Senecan tragedies (Fig. 2*A*, *i*). In contrast, total sense pauses were significantly reduced (*Octavia*) or enriched (*Hercules Oetaeus* and *Procne*) in the Senecan-influenced tragedies (Fig. 2*A*, *i* and *SI Appendix*, Fig. S1*A*, *i*), suggesting that the imitators either deliberately disregarded or failed to replicate a typical, if likely unconscious, aspect of Senecan style.

We then recapitulated a seminal literary critical study that used manual tabulation of sense-pause statistics to establish a relative chronology for the eight authentic tragedies (46). In contrast to total sense pauses, the ratio of intraline (sense pauses that do not coincide with line breaks in the iambic trimeter verse) to total sense pauses is more heterogeneous across the tragedies, as reported by Fitch (46) and supported by our computational analysis (Fig. 2*A*, *ii* and *SI Appendix*, Fig. S1*A*, *ii*). On the basis of this variation, Fitch (46) divided the tragedies into three groups, which we confirmed differ significantly in intraline to total sense-pause ratio (*SI Appendix*, Fig. S2). By analogy with the stylistic development of other playwrights, Fitch (46) further suggested that the ratio is higher in Seneca's later tragedies as the playwright became more skillful at exploiting tension between the basic units of meaning and meter. This relative chronology of Seneca's plays has been widely influential in classics, and even critics who disagree with Fitch's placements (46) of individual works have tended to retain the majority of his ordering (47). Fitch (46) excluded from his study the two tragedies in the corpus considered spurious. Ferri (48) has applied Fitch's method (46) to the *Octavia* but likewise used a manual count. In addition to rapidly confirming Fitch's three groupings (46), we also verified Ferri's discovery (48) that the *Octavia* has a relatively low ratio, similar to that expected for an early Senecan tragedy (Fig. 2*A*, *ii*). This result holds across multiple editions of Seneca, despite variations in absolute value of the ratio caused by differences in editorial practice (*SI Appendix*, Fig. S3). The stylistic resemblance of the *Octavia* to early Senecan tragedies is consistent with traditional critical assessments of the play as showing less technical virtuosity than most Senecan drama (48).

Enjambments are a special class of poetic sense pause, in which a sentence or clause "runs over" the end of a line of verse to the first word of the following line. We computationally tabulated enjambments in the tragedies by counting, in lines not starting with a new sentence, every punctuation mark (including commas) immediately after the first word. Counting punctuation is an effective heuristic for the identification of enjambments; for Correr's *Procne*, the precision was 0.97, and the recall was 1.0 (details are in *SI Appendix*, Text and Tables S1 and S2). Our analysis revealed a substantial (approximately threefold) enrichment of enjambments in Correr's *Procne* above any Senecan or classical pseudo-Senecan text (Fig. 2*A*, *iii* and *SI Appendix*, Fig. S1*A*, *iii*). As noted above, flexibility in the shape of the verse is typically considered as a mark of skillful poetic composition.

This variation stands in contrast to the monotony of an unbroken series of end-stopped lines (i.e., those lines in which the meaning is complete by the end of the line and marked by firm punctuation). One plausible explanation of the unusually high incidence of enjambment in the *Procne* is the desire of the young author—only 18 years old at the time—to display his virtuosity in Latin verse composition in part through the use of a feature that signified confident poetic technique. Although we possess no direct evidence of Correr's intent with respect to enjambment in particular, the playwright did preface his drama with a discussion of the varied meters used in the course of the text, including explicit discussion of meters that are rare in tragedy but more commonly found in comedies. Correr's frequent exploitation of enjambment can thus be considered complementary to his similar exploitation of the full array of Latin metrical forms, which went well beyond the range of meters used in Seneca's *Thyestes* (his primary classical model). The intertextual relationship between the *Procne* and its Senecan predecessors thus consists partly of similarities that highlight the tradition in which Correr is working and partly of differences (in this case, a difference in verse composition) that highlight Correr's distinctiveness within that tradition.

To investigate another potential stylistic difference, we next examined the use of relative clauses across the Senecan corpus. The relative clause, constructed using the relative pronoun who or which, is a standard method of subordinating one thought to another within a sentence. In Latin, relative pronouns are the various inflected forms of *qui* (*Materials and Methods* and *SI Appendix*, Text and Table S3 have details and error analysis). We computed the fraction of noninterrogative sentences with at least one relative clause for the 10 Senecan and pseudo-Senecan tragedies; interrogative sentences were excluded to obviate the need for semantic parsing of relative and interrogative pronouns, which are often identical morphologically. The count revealed that almost one-quarter of sentences in the *Octavia* contain a relative clause (Fig. 2*B* and *SI Appendix*, Fig. S1*B*), whereas the fraction for all other tragedies is below 20%. The *Octavia* stands out from the remainder of the corpus as a drama on a historical subject—the divorce and death of Nero's wife and the event's political context—in contrast to the mythological subjects of the other nine plays. The combination of non-Senecan authorship and historical subject matter has led critics to look for stylistic differences in the language and syntax of the work. With varying degrees of persuasiveness, claims have been made for the tragedy's comparatively less elaborate style, more colloquial speech, and features typically avoided in poetry (48). Our identification of the enrichment of relative clauses provides systematic, quantitative evidence that the *Octavia*'s syntax is distinctive from that of the other plays. The reason for this more hypotactic style is unclear. One possible explanation is that subordinating constructions of this kind indicate a more prosaic style, which could be an authorial habit or reflective of a more specific consideration. Partial corroboration of such a style can be found in specific instances identified by literary critics, such as the concatenation of relative clauses at lines 111 and 113 (48). The literary influence of Seneca's prose writing, especially the *De Clementia*, might also account for the *Octavia*'s more prosaic style (49).

Phonetic and Thematic Analyses of the *Octavia* and the *Phoenissae*. Functional n-grams are short, syllable-length strings of characters, which can reflect ingrained authorial style and capture patterns of sound in poetry. Analysis of functional n-grams has proven useful for authorship attribution studies and addressed literary questions in the postclassical reception history of the Roman poet Catullus (37). Although critics have long paid attention to specific aural effects and sound play in poetry, systematic studies have been infeasible without computational tabulation of n-grams.

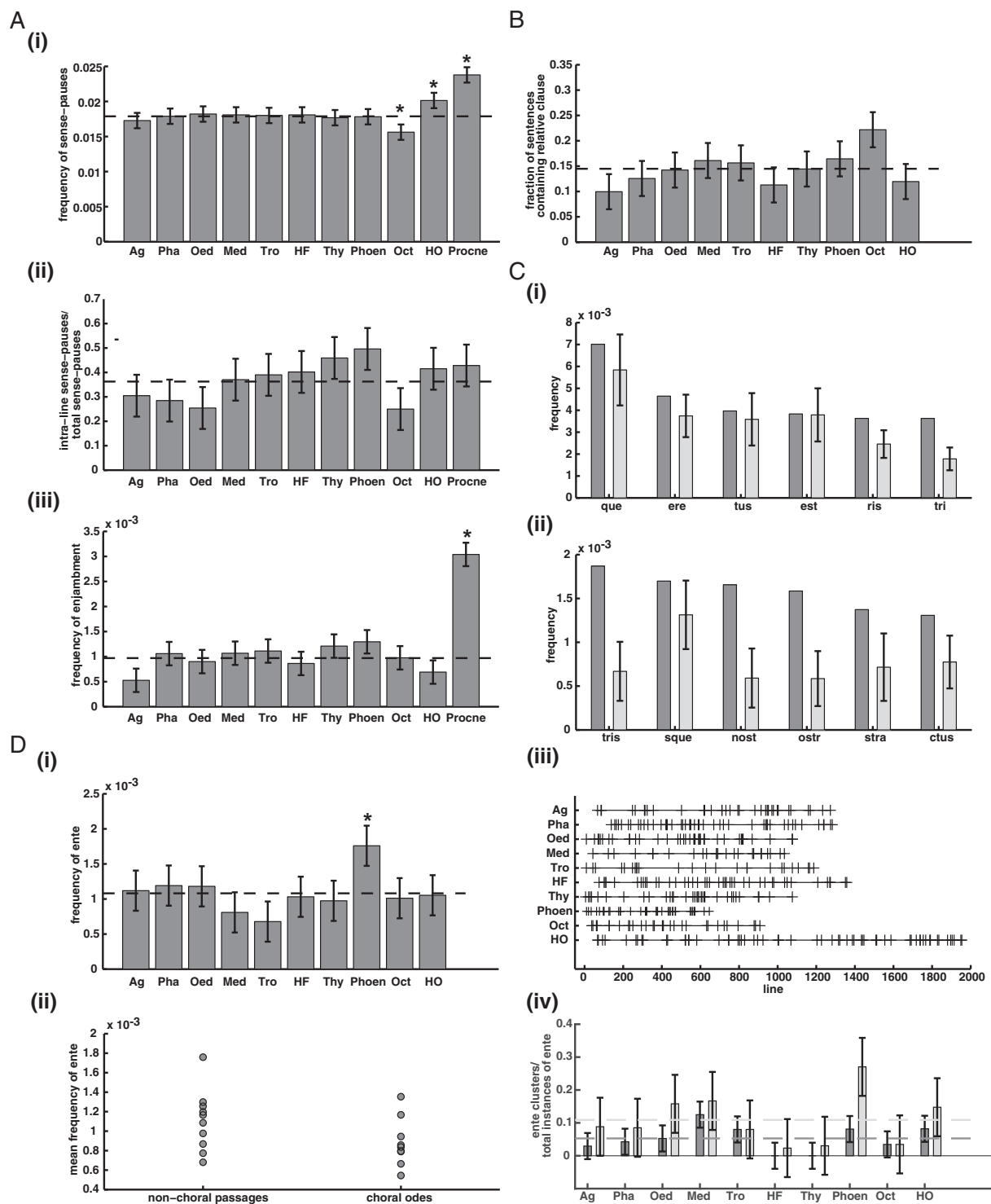


Fig. 2. Quantitative comparison of Senecan and pseudo-Senecan literary style. (A, i) Total sense pauses in each tragedy. (A, ii) Ratio of intraline to total sense pauses. (A, iii) Frequency of enjambment. (B) Fraction of noninterrogative sentences containing at least one relative clause. The *Octavia* is at Q3 + 1.46IQR, where Q is the quartile and IQR is the interquartile range. Frequencies of the five most common (C, i) three and (C, ii) four grams in the *Octavia* (dark gray bars). Light gray bars show the mean frequencies of each n-gram across the tragedies. (D, i) Frequency of the four-gram ente. (D, ii) Frequency of ente in choral and nonchoral passages. Each circle denotes the frequency in one tragedy. The *Phoenissae* lacks choral odes and was, therefore, excluded from the group on the right. The difference is nonsignificant ($p = 0.10$ by a two-tailed unpaired t test). (D, iii) Spatial distribution of ente in 10 tragedies. Each vertical line denotes one or more instances of ente at that position. (D, iv) Fraction of instances of ente that occur within clusters in each tragedy. The dark gray bars indicate instances within one line of each other, and the light gray bars indicate instances within three lines of each other. All frequencies are per character. In all plots, the dotted lines denote the mean of the relevant quantity across all tragedies, except the *Procne*. Error bars denote 1 SD across the tragedies. Senecan and pseudo-Senecan tragedies are referred to by abbreviations given in the Oxford Classical Dictionary: Ag, *Agamemnon*; HF, *Hercules Furens*; HO, *Hercules Oetaeus*; Med, *Medea*; Oct, *Octavia*; Oed, *Oedipus*; Pha, *Phaedra*; Phoen, *Phoenissae*; Tro, *Troades*; Thy, *Thyestes*. The *Procne* is a neo-Latin tragedy written in 1428 by Gregorio Correr. *Outliers (defined as $>Q3 + 1.5IQR$ or $<Q1 - 1.5IQR$).

We initially examined the most common functional bigrams (two-letter strings) in the *Octavia* and the *Hercules Oetaeus* and found that their frequency was comparable in both the spurious and authentic tragedies (*SI Appendix, Fig. S4*). This result prompted us to repeat the analysis for the *Octavia* with functional trigrams, for which we observed clear differences (Fig. 2C, *i*). Of particular interest, two of the six most common trigrams in the *Octavia* (tri and ris) are elevated compared with the authentic tragedies. The enrichment of particular n-grams points to the author's disposition toward a particular sound and possibly words containing those n-grams. In the case of the *Octavia*, those words are the various inflected forms of tristis (sad, stern) and noster (our), which together appear 69 times in the *Octavia* and account for more than 60% of the instances of tri and ris. The frequent use of tristis and noster is also reflected in the enrichment of the four-grams tris, nost, ostr, and stra (Fig. 2C, *ii*).

As an example of the kind of literary critical hypotheses that can be supported by analysis of functional n-grams, we might interpret the frequency of the appearance of tristis as substantiating the mood of lament and pessimism that pervades much of the *Octavia*, over and above what is typical even for Senecan tragedy. The enrichment of inflected forms of noster suggests a different but compatible hypothesis. Although the date and possible performance context of the *Octavia* are unknown, on the basis of its negative characterization of Nero scholars have argued that it was composed in the wake of Nero's death, either during or shortly after the period of civil wars known as the Year of the Four Emperors (AD 69). Much of the drama is concerned with Nero's tyrannical behavior and removal of opposition, and the play ends with mention of a popular uprising in support of Octavia. It thus dwells on various claims on political authority. The frequent use of the word noster (our) in the play repeatedly emphasizes the ownership that various parties feel over, for instance, the city (nostra urbs) or the imperial household (nostra domus). Resolving these rival claims is both the plot of the drama and a stimulus for the post-Neronian audience to reflect on the significance of such claims for their own time (discussed in detail in *SI Appendix, Text*).

Although written by Seneca, the *Phoenissae* has long been recognized as distinct from the remainder of the corpus (50). It is several hundred lines shorter than any other tragedy and obviously incomplete. Another distinctive aspect of the *Phoenissae* is that it does not contain any odes sung by a chorus, which are a standard component of Roman tragedy and present in all other Senecan and pseudo-Senecan tragedies. In our analysis of functional n-grams across the Senecan corpus, we found that the four-gram ente is significantly enriched in the *Phoenissae* (Fig. 2D, *i* and *SI Appendix, Fig. S1C, i*). This enrichment is specific to ente; related four grams, in which "nt" is immediately preceded and succeeded by any vowel, are not enriched in the *Phoenissae* (*SI Appendix, Fig. S5*). The enrichment of "vowel + nt + vowel" four grams in the *Thyestes* is a consequence of frequent references to Tantalus, an important character in that tragedy (*SI Appendix, Fig. S5*). Furthermore, there is no significant difference between the frequency of ente in choral and nonchoral passages across the Senecan corpus (Fig. 2D, *ii*), suggesting that the concentration of ente in the *Phoenissae* cannot be explained by its peculiar structure.

We examined the spatial distribution of instances of ente in the tragedies (Fig. 2D, *iii*), which revealed that the four gram is often repeated in close proximity in the *Phoenissae*. This effect, as measured by the fraction of instances of ente occurring within three-line clusters, is specific to the *Phoenissae* (Fig. 2D, *iv*). Additionally, clusters of the generic vowel + nt + vowel four gram are not enriched in any tragedy other than the *Thyestes* (*SI Appendix, Fig. S6*). As such, variations in its frequency might reflect some stylistic choice by the author, especially when clustered to create a partial echo.

Repetition of words for stylistic effect is a common feature of Senecan tragedy and the *Phoenissae* in particular, which exhibits frequent instances of exact repetition (e.g., sequor, sequor at 40 and ibo, ibo at 12 and 407) and morphological variation (e.g., patris ... pater at 55, frater ... fratrem at 355, and pectus ... pectori at 470). These formal repetitions often possess literary significance. In the *Phoenissae*, for instance, clusters of familial terms highlight the play's thematic focus on a civil war fought between two brothers (51). The repetitions cited by critics, however, operate at the level of the word (whether exact or a morphological variant) rather than purely phonetic elements, such as ente. Traditional critical approaches, based on reading or word searches, are thus poorly equipped to detect subtler forms of repetition manifested in smaller units.

The clusters of ente in the *Phoenissae* include repetitions of both whole words and morphological endings. Repetitions often serve to emphasize ideas or feelings important to the drama. At 368 and 369, for instance, Jocasta uses the word nocentes (guilty) in successive lines to amplify her sense of her own wrongdoing; n-gram analysis is especially useful for the identification of clusters of nonidentical, even etymologically unrelated words. To give one example, at 98–100, nolentem (unwilling) and cupientem (desiring) are paired in opposition to each other, a contrast highlighted by the aural echo of the ending. Other clusters of nonidentical words containing ente highlight themes of sexual aberration (467–469) and moral responsibility (451–454) that are important to the subject matter of the play (*SI Appendix, Text*).

Furthermore, we suggest that Seneca's greater propensity to exploit the repetition of this sound is consistent with the word-level repetitions already observed by critics as part of a larger stylistic aim. Seneca seems to use repeated words and sounds in close proximity in a systematic way. In dramatizing the mythological war between the twins Polynices and Eteocles, the *Phoenissae* is especially concerned with repetition, doubling, and assimilation—features that suffuse the speech, themes, and structure of the play. Although impossible to determine with any certainty, our inference about the frequent clustering of adjectival or participle endings in the *Phoenissae*, which are often used to signal apparent contrasts or amplifications, is that they embody at the level of sound a larger concern with repetition that defines the drama as a whole.

Anomaly Detection Differentiates Suspected Citations from Other Livian Material. We next considered citation and paraphrase, a class of intertextuality of comparable similarity but narrower scope than creative imitation of entire works (Fig. 14) and potentially amenable to techniques of authorship attribution. We took as a case study the use of source material in Livy's enormous history of Rome. The scope of Livy's writings required that he consult a wide variety of sources, mostly earlier historians but also published speeches and other texts. Like other historians, the manner in which Livy used his sources was equally varied, ranging from direct quotation and referential citation ("I found these numbers in X") to vague indications of a source ("some say," "I read somewhere") (45, 52, 53). Literary critics have also shown that, in certain places, Livy uses a specific source without explicitly saying so (54). The nature of Livy's source use is made even more opaque by the loss of most of the source texts in addition to the loss of the majority of his own history. Classical scholars have debated inconclusively the extent to which the text of earlier sources can be reconstructed from Livy's citational passages (i.e., passages that include a citational gesture, whether a reference to a specific author or a more indirect suggestion of source use) (55, 56). The paucity of extant source material poses an extreme challenge for standard stylometric identification (whether manual or computational) of Livian citations. Following our approach with pseudo-Senecan tragedy, we used a combination of computational and literary critical approaches

to achieve an improved understanding of Livy's citational practice. Our main result is the development of an anomaly detection algorithm that can differentiate Livian citations from noncitational material (i.e., the vast majority of the text) using stylistometric features.

Our analysis relied on a database previously developed by one of the authors (A.H.L.) for use in literary research, which catalogs citational passages in the extant parts of Livy's history. The database was compiled by noting all passages (in an English translation) in which Livy suggests use of source material, whether by explicit identification of a source or through citational language. In total, the database contains 439 citational passages.

We first performed a simple computational test to confirm the linguistic basis for the citation database. We compared the frequency of four representative citational phrases (*fama est*, it is rumored that; *annalibus*, in the annals; *scribit*, he writes; *tradit*, he reports) between the citation database and the rest of Livy and found, as expected, that these terms are enriched significantly in the database (Fig. 3*A*, *i*). We also examined the distribution of citations across Livy (Fig. 3*A*, *ii*). Over 50% of entries in the database occur in the first decade of Livy. Consistent with this enrichment of citations, the frequency of the citational phrase *annalibus* is significantly higher in the first decade (*SI Appendix*, Fig. S7).

We next assembled a large set of Latin stylistometric features that might be useful for distinguishing citational and noncitational material. The set consists of 25 features encompassing many items of stylistic interest, including noncontent words, specific syntactic constructions, and length of sentences and clauses (*SI Appendix*, Table S4). As discussed above, Livy's source texts are largely not extant, which precludes the application of binary classification. As an alternative, we used a one-class support vector machine (SVM) as an anomaly detection algorithm. The one-class SVM was trained on the Livian corpus (with some material excluded for cross-validation) and used to classify material in the citation database as anomalous (non-Livian) or nonanomalous (Livian). A primary challenge in the analysis of the citation database is the length of individual entries, many of which include only a few sentences. To generate meaningful feature statistics, we aggregated multiple citations into "bins" randomly and analyzed each bin as if it were a single passage (37). We set the bin size at 35 sentences, which was the minimum passage length for which we obtained consistent results (*SI Appendix*, Fig. S8). To maintain consistency, we also binned test material from Livy and other authors studied, even if extensive material was available.

For the citation database, we found that the fraction of bins classified as Livian was very low (less than 10%), regardless of the Livian material used for training (Fig. 3*B*). In contrast, ~80% of bins from Livian material withheld for cross-validation were classified as Livian. The correct identification of most of the cross-validation material as Livian and the substantial difference between the cross-validation material and the citation database validate the model as an effective tool for the analysis of citations. The fact that a small amount of Livian material was classified as anomalous likely reflects the well-known heterogeneity of Livy's style across 35 books of his history (57) and the general tendency of one-class anomaly detection methods to classify some test material as anomalous (58). For instance, Yilmazel et al. (59) used a one-class SVM to analyze a corpus of government documents and reported false negative rates between 29 and 47% (substantially higher than we obtained for Livy), depending on the features used.

We then investigated which of the stylistometric features were most effective for differentiating citational material. We reasoned that markers of hypotactic style (extensive use of subordinate clauses) might be particularly important, because the earlier

historians on whom Livy drew are generally held to have favored a simpler sentence structure (parataxis) in contrast to Livy's more varied and hierarchical syntax (60). Consistent with this hypothesis, we identified five features (mean sentence length, variance of sentence length, fraction of noninterrogative sentences containing at least one relative clause, mean length of relative clauses, and mean number of relative clauses per sentence) sufficient to establish a clear difference between citational and noncitational material (*SI Appendix*, Fig. S9). All five of these features relate to various aspects of the organization of sentences and together reflect tendencies toward hypotactic or paratactic style. Use of this low-dimensional feature set also enabled reduction of the bin size to 20 sentences (*SI Appendix*, Fig. S8) and a correspondingly finer-grained characterization of the citation database.

We applied our anomaly detection procedure with the reduced feature set to a passage that has provoked particular controversy over Livy's use of source material. Toward the end of Book 38, Livy describes a complicated sequence of events in the late career of Scipio Africanus, the famous Roman general. Focused primarily on the legal tribulations of Scipio and his brother, Livy's narrative is divided into two contrasting accounts, with the second largely undermining the first (61). The first account follows that of an earlier historian, Valerius Antias, whom Livy explicitly cites as a source. The second follows a number of other sources, including records of various speeches made by some of the principal participants in the events. Modern commentators have disagreed in particular on the extent to which Livy reused Valerius Antias, with judgments ranging from minimal reuse to extensive quotation (62). We applied our method to this narrative to ascertain whether there is a meaningful stylistic difference between the two accounts and determine which account, if either, differs from Livy's typical style. We divided the whole narrative into two sections large enough to include a substantial portion of text: the first (38.50.1–51.14) putatively more indebted to Valerius Antias, and the second (38.54.1–60.10) indebted to other sources. The one-class SVM classified the first section as "non-Livian" and the second section as "Livian." The result corroborates the view that Livy's first account was substantively influenced by Valerius Antias. However, it does not indicate whether such influence amounts to quotation, imitation, or a subtler stylistic effect. Both results have a shared implication for Livy specialists—that critical attention should focus less on the question of whether Livy quoted Antias and more on the question of the potential stylistic irregularities in the first account within the narrative.

Profiling the Development of Latin Prose Style. Given the clear difference observed between bulk Livy and the citation database, we next hypothesized that post-Livian historiography, and perhaps even imperial prose in general, would resemble bulk Livy more closely than citational material. The hypothesis was based on an assumption that Livy's sources would show traces of an earlier prose style, whereas Livy's own style was part of a more generally influential movement that would be reflected in later authors. Our approach was to assess the "Livianness" of 17 non-Livian texts using the reduced feature set and the same methodology applied to the citation database. We chose a wide-ranging corpus consisting of prose and poetry from a variety of genres and periods. The poetry was used as a control group. As expected, all five works—including comedy, tragedy, epic, and philosophical poetry from times before, after, and contemporaneous with Livy—scored as extremely non-Livian. The prose texts were also of various genres, including speeches, letters, and technical treatises in addition to historiography.

We observed a clear difference between most pre- and post-Livian prose. Of the pre-Livian material, the nonhistorical texts registered as very non-Livian, quite unlike Caesar's historiographical accounts of his wars in Gaul and a few years later

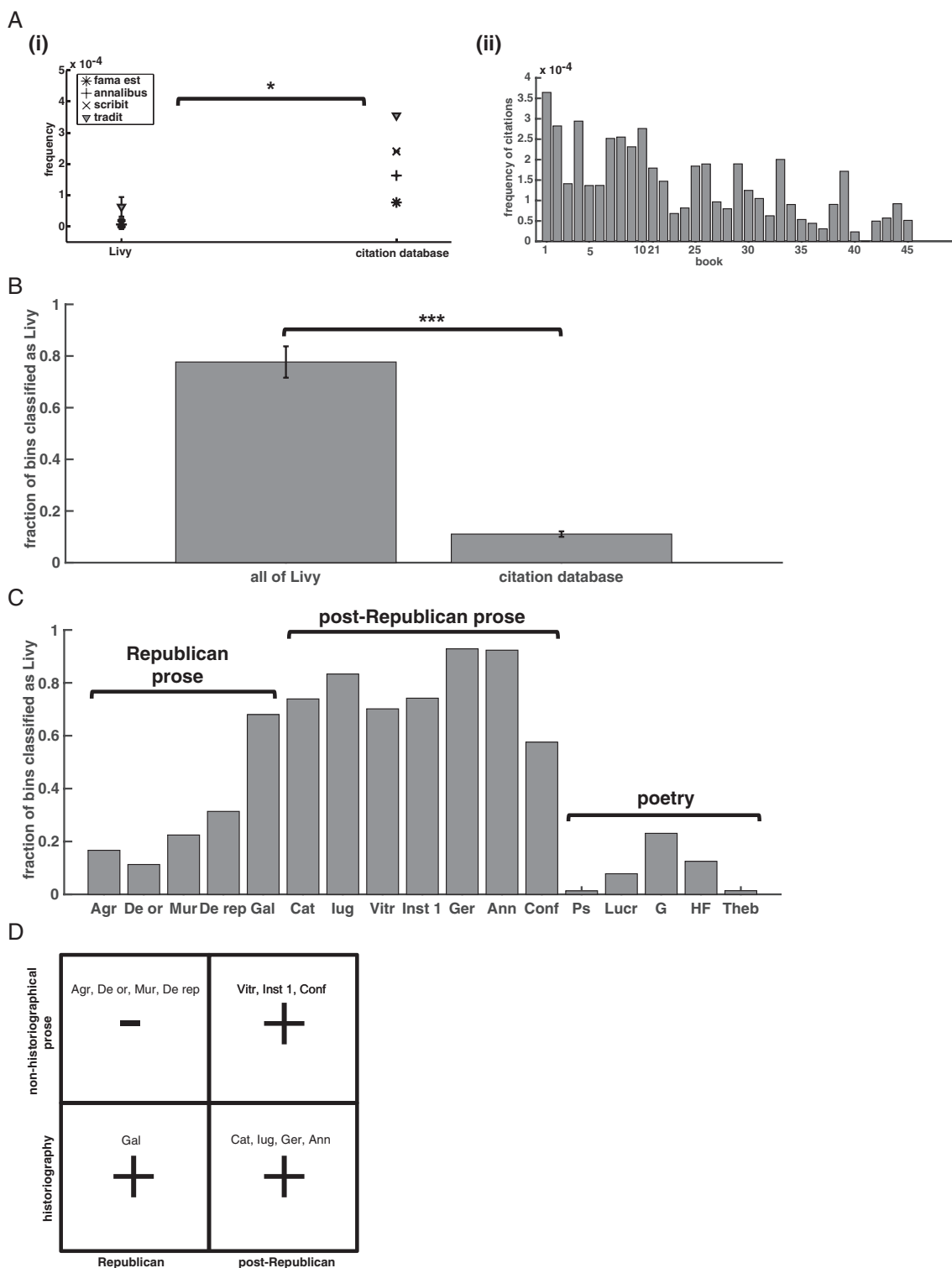


Fig. 3. Anomaly detection differentiates cited material from the rest of Livy. (A, i) Comparison of the frequency of four “signal words” indicating potential instances of citation (*fama est*, *annalibus*, *scribit*, and *tradiit*) between all of Livy (left) and the citation database (right). $*p < 0.05$ by a two-tailed unpaired *t* test. (A, ii) Frequency of entries in the citation database across 35 extant books of Livy. (B) Fraction of bins (random aggregates of 35 sentences) classified as Livian from bulk Livian material (left) and the citation database (right) by a one-class SVM using a set of 25 stylometric features. Results are the mean \pm 1 SD of 35 leave-one-out cross-validation experiments. $***p < 0.001$ by a two-tailed unpaired *t* test. (C) Fraction of 20-sentence bins from a range of Latin literature classified as Livian using a reduced set of five stylometric features. Works are referred to by abbreviations given in the Oxford Classical Dictionary: Agr, Cato’s *De Agri Cultura*; Ann, Tacitus’ *Annals*; Conf, Augustine’s *Confessions*; De or, Cicero’s *De oratore*; De rep, Cicero’s *De republica*; Cat, Sallust’s *De coniuratione Catilinae*; G, Vergil’s *Georgics*; Gal, Caesar’s *Bellum Gallicum*; Ger, Tacitus’ *Germania*; HF, Seneca’s *Hercules Furens*; Inst 1, Quintilian’s *Institutio Oratoria* 1; lug, Sallust’s *Bellum lugurthinum*; Lucr, Lucretius’ *De rerum natura*; Mur, Cicero’s *Pro Murena*; Ps, Plautus’ *Pseudolus*; Theb, Statius’ *Thebaid*; Vitr, Vitruvius’ *De architectura*. Genres represented include historiography (Gal, Cat, lug, Ger, and Ann), nonhistoriographical prose (Agr, De or, Mur, De rep, Vitr, Inst 1, and Conf), comedy (Ps), tragedy (HF), and poetry in dactylic hexameter (G, Lucr, and Theb). Prose and poetic texts are arranged chronologically. (D) Proposed outline of the development of Latin prose style; + indicates similarity to the style of Caesar and Livy.

Sallust's two monographs on historical topics, the *De coniuratione Catilinae* and the *Bellum Iugurthinum*. The result for Caesar's text, in particular, corroborates standard scholarly views about the resemblance between Caesar's and Livy's sentence structures and may reflect similarities in subject matter (57). The intermediate similarity of Cicero's *De re publica* suggests that content indeed plays a part in style. Unlike the two other Ciceronian works, a speech (*Pro Murena*) and a rhetorical treatise (*De oratore*), the *De re publica* contains more explicit discussions of history and politics in a narrative style. This fact may account for the work's greater resemblance to Livy's history. In the case of the later prose writers, however, even rhetorical (*Institutio Oratoria 1*) and technical (*De architectura*) treatises score as Livian, extending to Augustine's autobiographical *Confessions* written almost 400 years later. We note that two historiographical works by Tacitus (the *Germania* and the *Annales*) both seem particularly Livian in style (even slightly more so than bulk Livy). The difference between bulk Livy and Tacitus is far smaller than that between bulk Livy and the citation database or between early and later prose. The strong similarity, however, does suggest that Tacitus might have been influenced by Livy's syntax to a greater extent than has been appreciated previously (63).

On the whole, the two key observations are the difference between Livy and both pre-Livian prose and the material in the citation database and the similarity between Livy and Caesar and post-Livian prose. These results show in a quantitative and large-scale fashion a development in Latin prose style, namely that a stylistic shift occurred with Caesar, continued with Sallust and Livy, and exerted a critical influence on later prose literature (Fig. 3D). We find the effect of that influence even on genres, such as treatises, that had previously looked more unlike historiography. The results also reveal the extent to which Livy's citational material—whether in the form of imitations, quotations, or stylistic modulations—differs from later prose style.

Discussion

High-Throughput Data Generation for the Study of Literature and Culture. Numbers and statistics have long played an important, if underappreciated, role in literary criticism. Commentators often cite tabulations of particular words or formal features to bolster their arguments; in the mid-20th century, Duckworth (64) published a detailed quantitative study of meter in Latin poetry that, despite some issues of methodology, has had broad influence in the field of classics. In this regard, one obvious application of computation to literature is the replication, at larger scale and with greater efficiency, of standard stylistic studies. In our computational analysis of sense pauses in Senecan tragedy, we were able to both recapitulate Fitch's core results (46) efficiently and extend the scope of the original investigation. Accordingly, high-throughput methods are likely to have particular influence on the study of noncanonical material, such as the neo-Latin *Procne*, which receives negligible attention compared with famous classical authors, such as Vergil and Livy.

We find that frequency statistics on syllable-length n-grams can support literary criticism in two distinct but complementary ways. Highly enriched n-grams can point to patterns of word use that have thematic significance, as exemplified by our examination of *tristis* and *noster* in the *Octavia*. For such applications, the key advantage of functional n-gram analysis over simple word searches is that the former is untargeted, allowing for studies of diction even when the researcher does not have a specific hypothesis in mind. Additionally, functional n-grams enable the convenient investigation of colocalizations of sounds. Although criticism of poetry routinely reflects an intuitive understanding of aural effects, sound play and phonetic patterns are difficult to quantify using conventional methods. We suggest that analysis of short n-grams, an established technique in attribution studies and computational linguistics (65, 66), can inform literary critical

studies of poetry's aural quality. Functional n-grams are likely to be particularly useful when integrated with other computational approaches, such as the use by Forstall et al. (37) of functional bigrams as features for anomaly detection in literary texts.

Quantitative Criticism: Attribution, Interpretation, and the Digital Humanities. Computation has long been used for attribution and dating of literary works, problems that are unambiguous in scope and invite binary or numerical answers (27, 28). The recent explosion of interest in the digital humanities, however, has led to the key insight that similar computational methods can be repurposed to address questions of literary significance and style, which are often more ambiguous and open-ended. This turn from attribution to interpretation has been exemplified by the work of Jockers (10), who has pursued an approach to large-scale literary analysis termed "macroanalysis" (in analogy to macroeconomics). To this end, Jockers (10) has applied machine learning with stylistic features to trace patterns of influence across large English literary corpora, such as Victorian novels, and identify stylistic signatures of particular genres. Our analysis of the evolution of Latin prose style builds on such work in important ways. We repurpose anomaly detection to trace resemblances in a substantial corpus of Latin prose, identifying Caesar, Sallust, and Livy as a key point in the development of Latin prose style. These results suggest that later prose authors were influenced by the style of Caesar and the writers in Caesar's wake, including Livy, to a greater extent than has been previously acknowledged, even when writing about very different subject matter. Analogous phenomena have also been observed for the evolution of genres and literary style in English and other Latin corpora (7, 10, 25, 40). Throughout our work, we show the usefulness of incorporating syntactic and metrical features in addition to diction, noncontent words, and punctuation marks, which have been considered previously by Jockers (10) and others (25), into such comparative analyses.

Our approach, which we have termed "quantitative criticism," relies on a productive fusion of humanistic and computational methods. Although indebted to much groundbreaking work in the fields of computational text analysis and authorship attribution, we intend the reference to "criticism" to signal an equal debt to literary study's traditional concern with aesthetics and meaning. To that end, we seek to use quantitative data to understand literary relationships and literary interpretation to suggest quantitative experiments, so that the computational work of the scientist and the critical work of the humanist operate in symbiosis.

Materials and Methods

Editions of Texts. We used Peiper and Richter's 1921 edition of Seneca (67) and Weissenborn and Müller's 1911 edition of Livy (68) for all computational analyses. Both texts are freely and publicly available in searchable form through the Perseus Digital Library. For computational analysis of the *Procne*, we scanned Grund's 2011 text (69), applied optical character recognition, and manually corrected errors in the output. Sense-pause counts for the *Octavia* reported in *SI Appendix, Fig. S3* were determined manually using Giardina's 1966 text (70). All texts used in the comparison of Latin literary style reported in Fig. 3C are available through the Perseus Digital Library.

Computation of Stylistic Features. All natural language processing tasks were done using Python 2.7, and the code is freely and publicly available at <https://github.com/qcrit>. Copies of the relevant texts were obtained from the Perseus Digital Library as extensible markup language (XML) files and first stripped of all XML tags.

Following the definition of Fitch (46), sense-pause counts were determined by tabulation of punctuation marks other than commas [., ? , ! , ; , : , (,) , ~ , ' , ' , " , and "]. Enjambments were identified by noting instances of punctuation (including commas) that occurred after the first word of a line not immediately preceded by an end-line sense pause. A sentence was scored as having a relative clause if it was both noninterrogative (i.e., ending with a punctuation mark other than ?) and had at least one form of

the Latin relative pronoun (qui, cuius, cui, quem, quo, quae, quam, qua, quod, quorum, quibus, quos, quorum, or quas). We performed a manual error analysis of the procedures for enjambment and relative clause counting, which is discussed in *SI Appendix, Text and Tables S1–S3*.

For analysis of Livian citations, we considered a set of 25 stylistic features divided into five broad categories: pronouns, noncontent adjectives, conjunctions, subordinate clauses, and miscellaneous. The feature set is listed in *SI Appendix, Table S4*, and the methods used for calculating the features are described in *SI Appendix, Text*.

Assembly of Database of Possible Livian Citations. The database of Livian citations was constructed previously by one of the authors (A.H.L.). The method used to compile the database involved reading the entirety of Livy's history in English translation and noting all passages in which Livy names a source or uses citational language. Manual checks of portions of the Latin text found no instances of passages erroneously included. The database contains 439 distinct entries. The final corpus used for our analysis was created computationally by aggregating all passages of Livy mentioned in the database from the XML file of Weissenborn and Müller's text (68).

Anomaly Detection of Livian Citations. For anomaly detection, we used the scikit-learn implementation of a one-class SVM with a nonlinear (radial basis function) kernel and hyperparameters set to $\gamma = 1/25$ or $\gamma = 1/5$ (for the full and reduced feature sets, respectively) and $\nu = 1/5$ (71). As described

in the text, experiments were performed on randomly aggregated bins constructed from the texts analyzed. The bin size was determined empirically (*SI Appendix, Fig. S8*).

We trained the one-class SVM on the whole Livian corpus except for Book 1 using the full set of 25 stylistic features. We then classified all bins in the citation database and Book 1 as nonanomalous (Livian) or anomalous (non-Livian). This procedure was repeated 34 times, with one of the other extant books of Livy withheld for cross-validation each time. Fig. 3*B* reports mean fraction of bins classified as Livian over these 35 experiments. We then identified by direct experimentation a reduced set of five stylistic features for which we obtained comparable classifier performance (*SI Appendix, Fig. S9*). This reduced feature set was used for the analysis of Latin prose style reported in Fig. 3*C*.

ACKNOWLEDGMENTS. We thank Sarah Heiter for assistance with the error analysis of stylistic features and Krithika Iyer for help with natural language processing. We also thank Neil Coffee, Joe Farrell, Stephen Hinds, Dan Rockmore, and Ariane Schwartz for comments on the manuscript. This work was conducted under the auspices of the Quantitative Criticism Lab (www.qcrit.org), an interdisciplinary project codirected by J.P.D. and P.C. and supported by seed funding from the Office of the Provost at Dartmouth College, a Neukom Institute for Computational Science CompX Faculty Grant, and National Endowment for the Humanities Digital Humanities Start-Up Grant HD-248410-16. J.P.D. was supported by National Science Foundation Graduate Research Fellowship Grant DGE1144152, and P.C. was supported by an American Council of Learned Societies Digital Innovation Fellowship.

- Kristeva J (1980) Word, dialogue, and novel. *Desire and Language*, ed Roudiez LS (Columbia Univ Press, New York), pp 64–91.
- Juvan M (2009) *History and Poetics of Intertextuality* (Purdue Univ Press, West Lafayette, IN).
- Thomas RF (1986) Virgil's Georgics and the art of reference. *Harv Stud Class Philol* 90:171–198.
- Hinds SE (1998) *Allusion and Intertext: Dynamics of Appropriation in Roman Poetry* (Cambridge Univ Press, Cambridge, UK).
- Edmunds L (2001) *Intertextuality and the Reading of Roman Poetry* (Johns Hopkins Univ Press, Baltimore).
- Crane G (1996) Building a digital library: The Perseus Project as a case study in the humanities. *Proceedings of the First ACM International Conference on Digital Libraries* (Association for Computing Machinery, New York), Vol 1, pp 3–10.
- Moretti F (2005) *Graphs, Maps, Trees: Abstract Models for Literary History* (Verso, London).
- Moretti F (2013) *Distant Reading* (Verso, London).
- Ramsay S (2011) *Reading Machines: Toward an Algorithmic Criticism* (Univ of Illinois Press, Champaign, IL).
- Jockers M (2013) *Macroanalysis: Digital Methods and Literary History* (Univ of Illinois Press, Champaign, IL).
- Long H, So R (2016) Literary pattern recognition: Modernism between close reading and machine learning. *Crit Inq* 42:235–267.
- Hammond A, Brooke J, Hirst G (2013) A tale of two cultures: Bringing literary analysis and computational linguistics together. *Proceedings of the Second Workshop on Computational Linguistics for Literature* (Association for Computational Linguistics, Stroudsburg, PA), pp 1–8.
- Underwood T (2014) Theorizing research practices we forgot to theorize twenty years ago. *Representations* 127:64–72.
- Jockers ML, Mimmo D (2013) Significant themes in 19th-century literature. *Poetics* 41:750–769.
- Piper A (2015) Novel devotions: Conversational reading, computational modeling, and the modern novel. *New Lit Hist* 46:63–98.
- Bamman D, Crane G (2008) The logic and discovery of textual allusion. *Proceedings of the 2008 LREC Workshop on Language Technology for Cultural Heritage Data*. Available at citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.366.8213.
- Coffee N, Koenig JP, Porrima S, Ossewaarde R, Jacobson S (2012) Intertextuality in the digital age. *Trans Am Philol Assoc* 142:383–422.
- Coffee N, Koenig JP, Porrima S, Ossewaarde R, Jacobson S (2013) The Tesseract Project: Intertextual analysis of Latin poetry. *Lit Linguist Comput* 28:221–228.
- Bernstein N, Gervais K, Lin W (2015) Comparative rates of text reuse in classical Latin hexameter poetry. *Digit Humanit* Q 9(3).
- Coffee N, Bernstein N (2016) Digital methods and classical studies. *Digit Humanit* Q 10(2).
- Michel JB, et al. (2011) Quantitative analysis of culture using millions of digitized books. *Science* 331:176–182.
- Siebert R, Wellen C, Jin Y (2011) Spatial cyberinfrastructures, ontologies, and the humanities. *Proc Natl Acad Sci USA* 108:5504–5509.
- Aiden E, Michel JB (2013) *Uncharted: Big Data as a Lens on Human Culture* (Riverhead Books, New York).
- Lansdall-Welfare T, et al. (2017) Content analysis of 150 years of British periodicals. *Proc Natl Acad Sci USA* 114:E457–E465.
- Hughes JM, Fotia NJ, Krakauer DC, Rockmore DN (2012) Quantitative patterns of stylistic influence in the evolution of literature. *Proc Natl Acad Sci USA* 109:7682–7686.
- Lyu S, Rockmore D, Farid H (2004) A digital technique for art authentication. *Proc Natl Acad Sci USA* 101:17006–17010.
- Koppel M, Schler J, Argamon S (2009) Computational methods in authorship attribution. *J Am Soc Inf Sci Technol* 60:9–26.
- Stamatatos E (2009) A survey of modern authorship attribution methods. *J Am Soc Inf Sci Technol* 60:538–556.
- Hughes J, Graham D, Rockmore D (2010) Quantification of artistic style through sparse coding analysis in the drawings of Pieter Bruegel the Elder. *Proc Natl Acad Sci USA* 107:1279–1283.
- Stamou C (2008) Stylochronometry: Stylistic development, sequence of composition, and relative dating. *Lit Linguist Comput* 23:181–199.
- Mosteller F, Wallace DL (1964) *Inference and Disputed Authorship: The Federalist* (Addison-Wesley, Reading, MA).
- Holmes DI, Robertson M, Paez R (2001) Stephen Crane and the New-York Tribune: A case study in traditional and non-traditional authorship attribution. *Comput Humanit* 35:315–331.
- Vickers B (2004) *Shakespeare, Co-Author: A Historical Study of Five Collaborative Plays* (Oxford Univ Press, Oxford).
- Stover J, Winter Y, Koppel M, Kestemont M (2016) Computational authorship verification method attributes a new work to a major 2nd century African author. *J Assoc Inf Sci Technol* 67:239–242.
- Hoover DL (2007) Corpus stylistics, stylometry, and the styles of Henry James. *Style* 41:174–203.
- Hoover DL (2014) Modes of composition in Henry James: Dictation, style, and what Maisie knew. *Henry James Rev* 35:257–277.
- Bulson E (2014) Ulysses by numbers. *Representations* 127:1–32.
- Hope J, Witmore M (2010) The hundredth psalm to the tune of 'Green Sleeves': Digital approaches to the language of genre. *Shakespeare Q* 61:357–390.
- Eder M (2016) A bird's-eye view of early modern Latin: Distant reading, network analysis, and style. *Early Modern Studies After the Digital Turn*, eds Estill L, Jakacki D, Ulliyot M (Iter and ACMRS, Toronto), pp 63–89.
- Altmann E, Cristadoro G, Esposito MD (2012) On the origin of long-range correlations in texts. *Proc Natl Acad Sci USA* 109:11582–11587.
- Clement T, Tcheng D, Auviel L, Capitanu B, Monroe M (2013) Sounding for meaning: Using theories of knowledge representation to analyze aural patterns in texts. *Digit Humanit* Q 7(1).
- Levitan W (1989) Seneca in Racine. *Yale Fr Stud* 76:185–210.
- Miola RS (1992) *Shakespeare and Classical Tragedy: The Influence of Seneca* (Clarendon, Oxford).
- Haimson Lushkov A (2013) Citation and the dynamics of tradition in Livy's AUC. *Histos* 7:21–47.
- Fitch J (1981) Sense-pauses and relative dating in Seneca, Sophocles and Shakespeare. *Am J Philol* 102:289–307.
- Dingel J (2009) *Die Relative Datierung der Tragödien Senecas* (De Gruyter, Berlin).
- Ferri R (2003) *Octavia: A Play Attributed to Seneca* (Cambridge Univ Press, Cambridge, UK).
- Braund S (2009) *Seneca, De Clementia* (Oxford Univ Press, Oxford).
- Frank M (1995) *Seneca's Phoenissae: Introduction and Commentary* (Brill, Leiden, The Netherlands).
- Wills J (1996) *Repetition in Latin Poetry: Figures of Allusion* (Clarendon, Oxford).
- Fehling D (1989) *Herodotus and His 'Sources': Citation, Invention and Narrative Art* (Francis Cairns, Leeds, UK).

53. Grafton A (1997) *The Footnote: A Curious History* (Harvard Univ Press, Cambridge, MA).
54. Levene DS (2010) *Livy on the Hannibalic War* (Oxford Univ Press, Oxford).
55. Walsh PG (1961) *Livy: His Historical Aims and Methods* (Cambridge Univ Press, Cambridge, UK).
56. Forsythe G (1999) *Livy and Early Rome: A Study in Historical Method and Judgment* (Franz Steiner Verlag, Stuttgart).
57. Oakley S (1997) *A Commentary on Livy Books VI–X. Volume I, Introduction and Book VI* (Clarendon, Oxford).
58. Jain LP, Scheirer WJ, Boulton TE (2014) Multi-class open set recognition using probability of inclusion in computer vision. *Proceedings of the ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014*, eds Fleet D, Pajdla T, Schiele B, Tuytelaars T (Springer, Cham, Switzerland), Part III, pp 393–409.
59. Yilmaz O, Symonenko S, Balasubramanian N, Liddy ED (2005) Leveraging one-class SVM and semantic analysis to detect anomalous content. *Intelligence and Security Informatics. ISI 2005. Lecture Notes in Computer Science*, eds Kantor P, Muresan G, Roberts F, Zeng D, Wang F-Y, Chen H, Merkle R (Springer, Berlin), Vol 3495, pp 381–388.
60. Briscoe J (2005) The language and style of the fragmentary republican historians. *Aspects of the Language of Latin Prose*, eds Reinhardt T, Lapidge M, Adams J (Oxford Univ Press, Oxford), pp 53–72.
61. Haimson Lushkov A (2010) Intertextuality and source criticism in the Scipionic trials. *Livy and Intertextuality*, ed Polleichtner W (Wissenschaftlicher Verlag Trier, Trier, Germany), pp 93–133.
62. Briscoe J (2008) *A Commentary on Livy Books 38–40* (Oxford Univ Press, Oxford).
63. Oakley S (2009) Style and language. *Cambridge Companion to Tacitus*, ed Woodman A (Cambridge Univ Press, Cambridge, UK), pp 195–211.
64. Duckworth GE (1969) *Vergil and Classical Hexameter Poetry: A Study in Metrical Variety* (Univ of Michigan Press, Ann Arbor, MI).
65. Cavnar W, Trenkle J (1994) N-gram based text categorization. *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval*. Available at citeseerx.ist.psu.edu/viewdoc/similar?doi=10.1.1.53.9367&type=ab.
66. Houvardas J, Stamatatos E (2006) N-gram feature selection for authorship identification. *Proceedings of Artificial Intelligence: Methodology, Systems, and Applications (AIMSA)*, eds Euzennat J, Domingue J (Springer, Berlin), pp 77–86.
67. Peiper R, Richter G (1921) *L. Annaei Senecae Tragoediae* (Teubner, Leipzig, Germany).
68. Weissenborn W, Müller HJ (1880–1911) *Titi Livi ab urbe condita libri* (Weidmann, Berlin).
69. Grund GR (2011) *Humanist Tragedies* (Harvard Univ Press, Cambridge, MA).
70. Giardina GC (1966) *L. Annaei Senecae Tragoediae* (Editrice Compositori, Bologna, Italy).
71. Pedregosa F, et al. (2011) Scikit-learn: Machine learning in Python. *J Mach Learn Res* 12:2825–2830.

Quantitative criticism of literary relationships

Joseph P. Dexter,^{1,2} Theodore Katz,¹ Nilesh Tripuraneni,¹ Tathagata Dasgupta,¹ Ajay Kannan, James A. Brofos, Jorge A. Bonilla Lopez, Lea A. Schroeder, Adriana Casarez, Maxim Rabinovich, Ayelet Haimson Lushkov, and Pramit Chaudhuri²

SUPPORTING INFORMATION APPENDIX

SI Appendix, Text

This text has three primary objectives. It discusses validation of the heuristics used for computation of some of the stylometric features (*SI Appendix*, Tables S1-S3), it provides a more detailed literary critical interpretation of the Seneca data (Fig. 2; *SI Appendix*, Figs. S1-S6), and it describes the full set of features used for analysis of Livian citations (*SI Appendix*, Table S4). It should be read in conjunction with the *Results* section of the main paper.

Error analysis of enjambment calculations. The computational identification of enjambments relied on punctuation. As described in the *Materials and Methods* section of the main paper, we counted any sense-pause (including commas) that occurred after the first word of a line as an enjambment unless there was also a sense-pause at the end of the previous line. However, punctuation after the first word in a verse line occasionally is used not to mark a sense-pause of any literary significance, but rather to set off a subsequent address to a named individual or entity (in grammatical terms, the name typically appears in the vocative case). We manually tabulated enjambments in two sample plays, Seneca's *Phoenissae* and Correr's *Procne*, and compared the results with the computational tallies. We found no instances of false negatives (i.e., true enjambments missed by the punctuation counting procedure) and a small number of false positives, all but one of which involved a vocative at the beginning of the line. We counted 27 true enjambments and three false positives for the *Phoenissae*, and 89 true enjambments and three false positives for the *Procne*. As such, the precision of the enjambment heuristic is 0.9 and the recall 1.0 for the *Phoenissae*; for the *Procne*, 0.97 and 1.0. Sentences containing misidentified enjambments are listed in *SI Appendix*, Tables S1 (*Phoenissae*) and S2 (*Procne*).

Enjambment in Correr's *Procne*. There are numerous examples of Correr's sensitivity to the attention-grabbing effects made possible by enjambment. In an address to the god Mars, for instance, the character Tereus refers to himself in an enjambed line: *inclitum cernis, pater, / gnatum*. ("you, father, behold your famous son," *Procne* 142-143). Tereus thus draws attention to his divine birth and his relationship to Mars through the enjambment, which places emphasis on the word "son" (*gnatum*) occurring immediately after "father" (*pater*) and yet on the next line, marked by a firm pause (the period following *gnatum*). The arrangement of words makes adjacent and yet separates two familial terms that intuitively belong together in a way that cannot easily be replicated in English translation. Correr's interest in the relationship between Tereus and Mars, highlighted here in the disposition of the words "father" and "son," is corroborated by the preface to the play, which explicitly mentions the mythical genealogy linking the two figures.

The striking frequency of enjambment in the *Procne* compared with Senecan and pseudo-Senecan tragedy of the classical period may point to further, and necessarily more speculative, literary critical hypotheses. Both of Correr's classical models, Ovid's account of the myth in the *Metamorphoses* and especially Seneca's *Thyestes*, are explicitly concerned with the idea of surpassing one's predecessors and of excessiveness in general. It is possible, then, that the preponderance of enjambment in the *Procne* reflects Correr's youthful exuberance to outdo his classical forebears in the context of a play that itself thematizes oneupmanship. On this view, Correr's frequent use of enjambment has semantic as well as stylistic value: through its repeated deployment, the technique evokes the idea of exceeding a limit (represented by the end of the verse line), which in turn reflects the thematic concerns of the play and its prior tradition. Although it is impossible to prove the interpretation, the example nevertheless illustrates the productive combination of quantitative and literary critical approaches. The rapid computational calculation of a standard poetic feature such as enjambment can lead directly to the generation of interesting, albeit speculative, literary critical hypotheses.

Error analysis of relative clause calculations. Latin relative pronouns and interrogative pronouns/adjectives/adverbs have very similar forms. For instance, *quem* can mean either "whom" (relative pronoun), "whom?" (interrogative pronoun), or "which [person or thing]?" (interrogative adjective) depending on the syntax of the sentence. Our aim was to investigate complex subordination of sentences (indicated by relative pronouns) as a marker of authorial style. This goal entailed computationally counting instances of relative pronouns, but not interrogative pronouns or adjectives, without recourse to semantic parsing. Our approach, described in the *Materials and Methods* section of the main paper, was to exclude all direct interrogative sentences (i.e., those ending in a question mark), since interrogative sentences are much more likely than non-interrogative sentences to contain an interrogative pronoun that could be misidentified as a relative pronoun. We performed a manual error analysis of our relative pronoun counts using a sample corpus that consisted of two tragedies (*Phoenissae*, *Octavia*) and a quarter of one book of Livy (22.1-15).

¹J.P.D., T.K., N.T., and T.D. contributed equally to this work.

²To whom correspondence may be addressed. Email: jdexter@fas.harvard.edu or pramit.chaudhuri@austin.utexas.edu.

We first checked indirect interrogative sentences (questions reported by the author or a speaker rather than being posed directly, which therefore do not end in a question mark and were not excluded) for instances of interrogative pronouns and adjectives (i.e., false positives). Our manual tabulation found no instances of an interrogative pronoun or adjective within an indirect question in the *Phoenissae* and *Octavia*, and only two instances in the sample of Livy (*SI Appendix*, Table S3). We then checked for instances of relative pronouns within direct interrogative sentences (i.e., false negatives). The number of false negatives exceeds the number of false positives, but remains low compared with the total number of relative clauses in non-interrogative sentences (*SI Appendix*, Table S3). As reported in *SI Appendix*, Table S3, the precision for our heuristic ranged from 0.97 to 1.0 depending on the text examined, and the recall from 0.77 to 0.88. The analysis of the sample texts therefore suggests that the method is sufficient to support our inferences regarding syntactical style in Seneca, Livy, and other Latin authors.

Instances of the adverb *quam* (typically meaning “than” in comparisons or “how” in questions) or of the conjunction *quod* (meaning “because”) are also likely to have been miscounted as an identical form of the relative pronoun. However, such uses are considerably less frequent than the relative pronoun and hence are unlikely to have a substantial impact on the calculated relative clause frequencies.

Diction, style, and theme in the *Octavia*. As described in the main text, our analysis of functional n-grams in the *Octavia* identified two words both frequent in and thematically important for the play, *noster* (“our”) and *tristis* (“sad,” “stern”). The main objectives of this supplementary discussion are to cite additional literary evidence in support of our analysis, and to elaborate on the implications of our findings for understanding the themes of the drama.

First- and second-person possessive pronouns (*meus*, “my;” *tuus*, “your”) are unusually common in the *Octavia*. A longstanding argument explains the prevalence of such words in terms of versification and compositional style rather than semantic significance (1). On this view, the poet takes over a reasonably common Ovidian and Senecan disyllabic line-ending and uses it excessively. This habit contributes to a more general critique of the competent though not outstanding abilities of the poet, who is able to follow Senecan style but falls short of his exemplar’s level.

The first-person plural possessive *noster* (“our”), already highlighted as an important term using our functional n-gram analysis, is not deployed by the poet in the same way as *meus*, *tuus*, and other disyllabic possessives (e.g., *suus*, “his/her/its own”). The overwhelming majority of instances of the latter words and their grammatical inflections appear at line-end (*meus*: 44 line-end / 10 mid-line, *tuus*: 30 / 12, *suus*: 32 / 6). In marked contrast, *noster*, which differs prosodically from *meus* and *tuus*, appears far more commonly mid-line, with almost no line-end examples (3 line-end / 39 mid-line). In other words, whatever motivates the poet to use *noster* with great frequency, it is not the same habit of versification that plausibly underlies the placement of other possessives.

Even if a large proportion of the possessive pronouns are best explained as the product of the poet’s versifying tendencies, their collective prevalence bears on the themes of the drama. The plot of the *Octavia* concerns the divorce and exile of the emperor Nero’s wife (the eponymous Octavia), Nero’s marriage to his mistress Poppaea, and the tyrannical excesses of his character. On a literary analysis, possessive pronouns - especially first-person (*noster*, *meus*) and second-person (*tuus*) pronouns - are directly connotative of ownership and suggestive of a personal perspective on events. The *Octavia* is a play in which rival claims to possession are perhaps more central, and are certainly more numerous, than in other Senecan tragedies: the first wife vs. the second, Nero vs. the stepbrother he has murdered, Nero vs. his political advisor Seneca (who appears as a character within the drama), Nero vs. the chorus of Roman people (who favor Octavia), to mention only the largest contentions. In addition, there are multiple struggles over the sites that various parties lay claim to: the city, the household, the bedroom.

The combination of ownership and personal perspective takes on an especially political coloring in several of the phrases in which *noster* appears. Consider the following words used with *noster*, with the speaker or speakers noted in parentheses: *domus* (“household;” Octavia, Nero), *princeps* (“emperor;” Chorus, Octavia), *dux* (“leader;” Chorus), *urbs* (“city;” Nero, Chorus), *saeculum* (“age;” Nero). In each case *noster* is attached to a political or politicized entity, whether the imperial household, the emperor himself, the city, or even the age defined by Nero’s reign. In some cases the word is used as a genuine plural (e.g., by the chorus), in other cases as a royal “our” (e.g., by Nero). But beyond such linguistic parsing of *noster* lies a prior and more important question: whether these entities should be seen as belonging to one person or another, or even to a group. This question of ownership drives the struggle between members of the imperial household and, at a larger scale, between tyrant and people. Nero was notorious for treating (and mistreating) as his own what should belong to others or to a wider constituency (cf. Tacitus, *Annales* 15.45.1). This attitude is precisely characteristic of tyranny, and the critique of it is highly appropriate subject matter for a follower of Seneca writing some years in the wake of Nero’s fall.

Although traditional scholarship attributes the frequency of possessives to the poet’s crude versification, a combination of n-gram analysis (which highlighted *noster*) and philological study (which highlighted several possessive pronouns as a class) led to alternative hypotheses about the importance of such words. These hypotheses were in turn corroborated and fleshed out in a qualitative fashion using the techniques of literary criticism. The author of the *Octavia* may have been a more formulaic poet than Ovid or Seneca, but a more charitable interpretation of his diction is enabled by the use of quantitative analysis applied in tandem with traditional critical practices.

Our attention to possessive pronouns also has a bearing on interpretation of the adjective *tristis* (“sad” or “stern”), the other word besides *noster* highlighted by the n-gram analysis as being especially enriched in the *Octavia*. Based on the n-gram analysis alone, we postulated that the word’s frequency might create a mood of melancholy, lament, or suffering. That notion appears to find orthogonal support from other aspects of the play’s diction. In surveying Octavia’s uses of *meus*, we observe that many instances refer to her *fortuna* (“fortune”), *casus* (“misfortune”), *mala* (“evils”), *luctus* (“grief”), and *fata* (“fate”). These

moments of unhappy self-reflection bolster our claim about the heightened mood of lament due to the frequent appearances of *tristis*. These various expressions attribute an unusually pessimistic cast to the *Octavia*, even in comparison to a Senecan corpus generally characterized by harshness and gloom.

Phonetic clustering in the *Phoenissae*. Three examples of clusters of “ente” four-grams in the *Phoenissae* illustrate the potential literary significance of this anomalous feature within the Senecan corpus.

Significant repetitions need not be adjacent. “Ente” clusters at one- or two-line intervals are especially enriched in the *Phoenissae* compared with the rest of the corpus. It may seem counterintuitive, especially for readers accustomed to poetry characterized by rhyming endings of successive or alternating verse lines (as in much English poetry), that a writer might exploit echoes of sound at greater intervals. *Phoen.* 314-319, which contains a triple repetition of the phrase *iubente te* (“if you give the order”) at the beginning of the verse line, illustrates Seneca’s exploitation of sound echoes both in adjacent lines (318-319) and at greater intervals (314): *iubente te . . . / iubente te, praebebit alitibus iecur; / iubente te, vel vivet* (“if you give the order . . . / if you give the order, he will offer his liver to the birds; if you give the order, he will even live”). Although 318-319 contain an adjacent repetition, the first instance of *iubente te* occurs several lines earlier at 314. The effect of the word arrangement is to shorten the period of repetition, first felt at 318 as a distant echo of the initial phrase four lines earlier, only to become closer and more emphatic with the third occurrence in the immediately following line, which is the climax and culmination of Oedipus’ speech.

Significant repetitions need not be restricted to whole words. Perhaps the most striking instance of a repetition of “ente” occurs when Jocasta urges her exiled son Polynices to put down his weapons and end the siege of his home city, Thebes. In the context of a play about the effects of an incestuous marriage, a play that literary critics have often identified as sexually suggestive, Jocasta uses perhaps the most jarring innuendo in Latin literature: *claude vagina impium / ensem, et trementem iamque cupientem excuti / hastam solo defige* (“Sheathe your impious sword in its scabbard, and plant your trembling spear, which already desires to be cast down, in the ground,” *Phoen.* 467-469) (2). The language of sheathes, weapons, and desire leaves almost no room for ambiguity, and in this already erotically charged context it may even be that the audience is supposed to hear in the sound of the word *trementem* an allusion to the Latin word for penis, *mentula* (3). With specific regard to the repetition of “ente,” the jingle of participle endings would here seem to draw further emphasis to the psychological push and pull (“trembling” and “desiring”) characteristic of this most Freudian of dramas.

Significant repetitions can be both non-adjacent and not restricted to whole words. Our third and final example, though less spectacular than the previous one, best encapsulates the interest of the four-gram data (*Phoen.* 451-454):

error invitos adhuc
fecit nocentes: omne Fortunae fuit
peccantis in nos crimen: hoc primum nefas
inter scientes geritur.

Error has made me, though unwilling,
nonetheless guilty: the crime was all Fortune’s,
doing us wrong: this is the first sin
committed knowingly.

Here we see non-adjacent clustering of non-identical words that share the same morphological ending. The meaning of the clauses is contrasted but the words themselves are not antonyms, as are *nolemtem* (“unwilling”) and *cupientem* (“desiring”) at 98-100. It is in part the similar sound of the two words *nocentes* and *scientes*, perhaps augmented by *peccantis*, which reinforces the comparison, and ultimately the opposition, between the two clauses. Here is a simple yet effective instance of local sound repetition - identified computationally - contributing to the structure and semantics of a passage.

Computation of stylometric features. We computed a set of 25 Latin stylometric features for use in the anomaly detection experiments, which was subsequently narrowed to a reduced set of five features. All features are continuous, were computed without use of syntactic parsing, and fall into five broad categories (*SI Appendix*, Table S4). The features in the first two categories (pronouns and non-content adjectives) were calculated by counting instances of the various inflected forms of the indicated Latin word(s). Tables of the inflected forms can be found in any standard textbook or reference grammar for Latin, such as *Allen and Greenough’s New Latin Grammar* (freely available through the Perseus Project at <http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.04.0001&redirect=true>).

Some features comprised whole words, others comprised sequences of characters within words. For example, if counting instances of the polysemous word *ut*, which is both an adverb and a conjunction, we computed all appearances of the n-gram as a single word (e.g., *ut geniti*, *ut educati*, *ut cogniti essent*, not *Turnus rex Rutulorum*.) When counting morphological forms such as superlative endings, however, we computed all instances of the relevant n-gram within a word (e.g., *opulentissima*, where the n-gram *-issim-* is common to all standard superlative endings). All frequencies in the feature set are per-word.

We selected a diverse range of grammatical and syntactical categories to increase the chance of capturing stylistic patterns of different kinds. Although some features could be calculated with perfect accuracy (e.g., counts of n-grams), without the aid of syntactic parsing other features could only be approximated using heuristics. Error analysis was performed for a small

sample of these features (*SI Appendix*, Table S3). In general, the accuracy or comprehensiveness of the feature counts is not uniform, and some features were chosen with the understanding that only a small subset of instances were being counted (e.g., gerunds and gerundives).

Conjunctions:

- Conjunctions were computed by counting all instances of *et*, *-que*, *atque*, *ac*, *neque*, *aut*, *vel*, *at*, *autem*, *sed*, *tamen*, *postquam*.
- Frequency of *atque* followed by a consonant was computed by counting all instances of *atque* immediately followed by a word that begins with a consonant.

Subordinate clauses:

- Conditional clauses were computed by counting all instances of the words *si*, *nisi*, *quodsi*.
- *cum* clauses (where *cum* is an adverb or conjunction, but not a preposition) were computed by counting all instances of *cum* that are not immediately followed by a word ending in: *-a*, *-is*, *-e*, *-ibus*, *-ebus*. The limitations were applied to exclude instances of *cum* as a preposition (which is followed by nouns in the ablative case, several inflected endings of which are listed above).
- *quin* clauses were computed by counting all instances of *quin*.
- *antequam* clauses were computed by counting all instances of *antequam*.
- *priusquam* clauses were computed by counting all instances of *priusquam*.
- *dum* clauses were computed by counting all instances of *dum*.
- The fraction of non-interrogative sentences containing at least one relative clause was calculated as follows: a sentence was scored as having a relative clause if it was both non-interrogative (i.e., ending with a punctuation mark other than “?”) and had at least one form of the Latin relative pronoun (*qui*, *cuius*, *cui*, *quem*, *quo*, *quae*, *quam*, *qua*, *quod*, *quorum*, *quibus*, *quos*, *quarum*, or *quas*). Interrogative sentences were excluded to obviate the need for semantic parsing of relative and interrogative pronouns, which are often identical morphologically.
- The mean length of relative clauses was calculated by counting the number of characters in relative clauses identified as above.
- The number of relative pronouns per non-interrogative sentence was calculated by dividing the total number of relative pronouns in non-interrogative sentences by the total number of non-interrogative sentences. Interrogative sentences were excluded for the reasons given above.

Miscellaneous:

- (Direct) interrogative sentences were computed by counting all instances of a sentence ending in a question mark.
- Standard superlative adjectives and adverbs were computed by counting all instances of *-issim-* within a word. The method excluded certain common superlatives such as *maximus* or *optimus*, which would be difficult to capture precisely without also incorporating proper names (e.g., Fabius Maximus, Jupiter Optimus Maximus).
- *ut* clauses (where *ut* is an adverb or a conjunction) were computed by counting all instances of *ut*.
- The limited subset of gerunds and gerundives was computed by counting all instances of *-ndus* and *-ndum*. The restriction was designed to exclude the many verb forms that share the same letter sequence as the characteristic gerundival ending (e.g., *defendo*, *pendo*), though at the cost of also excluding the majority of the inflected forms of the gerund and gerundive. Erroneous inclusion of adjectives of the form *blandus* were assumed not to vitiate the count.
- The mean length of sentences was calculated by counting the number of characters in sentences ending in a “.” “?” or “!” and computing the mean. We excluded from the count any periods occurring after a single standalone character, since such instances typically indicate an abbreviation of a proper name rather than a sentence-end.
- Sentence length variance was calculated by counting the number of characters in sentences ending in a “.” “?” or “!” and computing the variance. We excluded from the count any periods occurring after a single standalone character for the reason given above.

1. Ferri R (2003) *Octavia: A Play Attributed to Seneca*. (Cambridge Univ Press, Cambridge, UK).

2. Ginsberg L (2015) Don't stand so close to me: Antigone's *pietas* in Seneca's *Phoenissae*. *Trans Am Philol Assoc* 145:199–230.

3. Adams J (1982) *The Latin Sexual Vocabulary*. (Johns Hopkins Univ Press, Baltimore, MD).

SI Appendix, Tables

Reference	Misidentified Enjambment
74-75	<i>non deprecor, non hortor, extingui cupis votumque, genitor, maximum mors est tibi?</i>
232-233	<i>... et aures ingerunt quicquid mihi donastis, oculi, cur caput tenebris grave</i>
520-521	<i>quantum daturus: 'quando pro te desinam' dixi 'timere?' dixit inridens deus:</i>

Table S1. Specific instances of misidentified enjambments in Seneca's *Phoenissae*.

Reference	Misidentified Enjambment
517-518	<i>Bacchis lampade nos vocat</i> <i>Euboe, Oggigie, adveni!</i>
542-543	<i>Mundus certa decentia</i> <i>munus, Bacche, tuum tulit.</i>
751-752	<i>Disce ex marito denique insigne facinus</i> <i>audere, Progne!</i>

Table S2. Specific instances of misidentified enjambments in Correr's *Procne*.

	TP	FP	FN	Precision	Recall
<i>Octavia</i>	77	0	14	1.0	0.85
<i>Phoenissae</i>	43	0	13	1.0	0.77
Livy 22.1-15	67	2	9	0.97	0.88

Table S3. Error analysis of relative clause frequency. The table lists the true positives, false positives, false negatives, precision, and recall for identification of relative clauses in the three sample texts.

	pronouns
1	frequency of personal pronouns
2	frequency of demonstrative pronouns
3	frequency of <i>quidam</i>
4	frequency of third-person reflexive pronouns
5	frequency of <i>iste</i>
	non-content adjectives
6	frequency of <i>alius</i>
7	frequency of <i>ipse</i>
8	frequency of <i>idem</i>
	conjunctions
9	aggregate frequency of conjunctions
10	frequency of <i>atque</i> followed by a consonant
	subordinate clauses
11	frequency of conditional clauses
12	frequency of <i>cum</i> clauses
13	frequency of <i>quin</i> clauses
14	frequency of <i>antequam</i> clauses
15	frequency of <i>priusquam</i> clauses
16	frequency of <i>dum</i> clauses
17	fraction of sentences containing a relative clause
18	mean length of relative clauses
19	number of relative clauses per sentence
	miscellaneous
20	frequency of interrogative sentences
21	frequency of superlatives
22	frequency of <i>ut</i> clauses
23	frequency of selected gerunds and gerundives
24	mean sentence length
25	variance of sentence length

Table S4. Full feature set for stylistic analysis of Livian citation. The 25 features are divided into five broad grammatical and syntactical categories.

SI Appendix, Figures

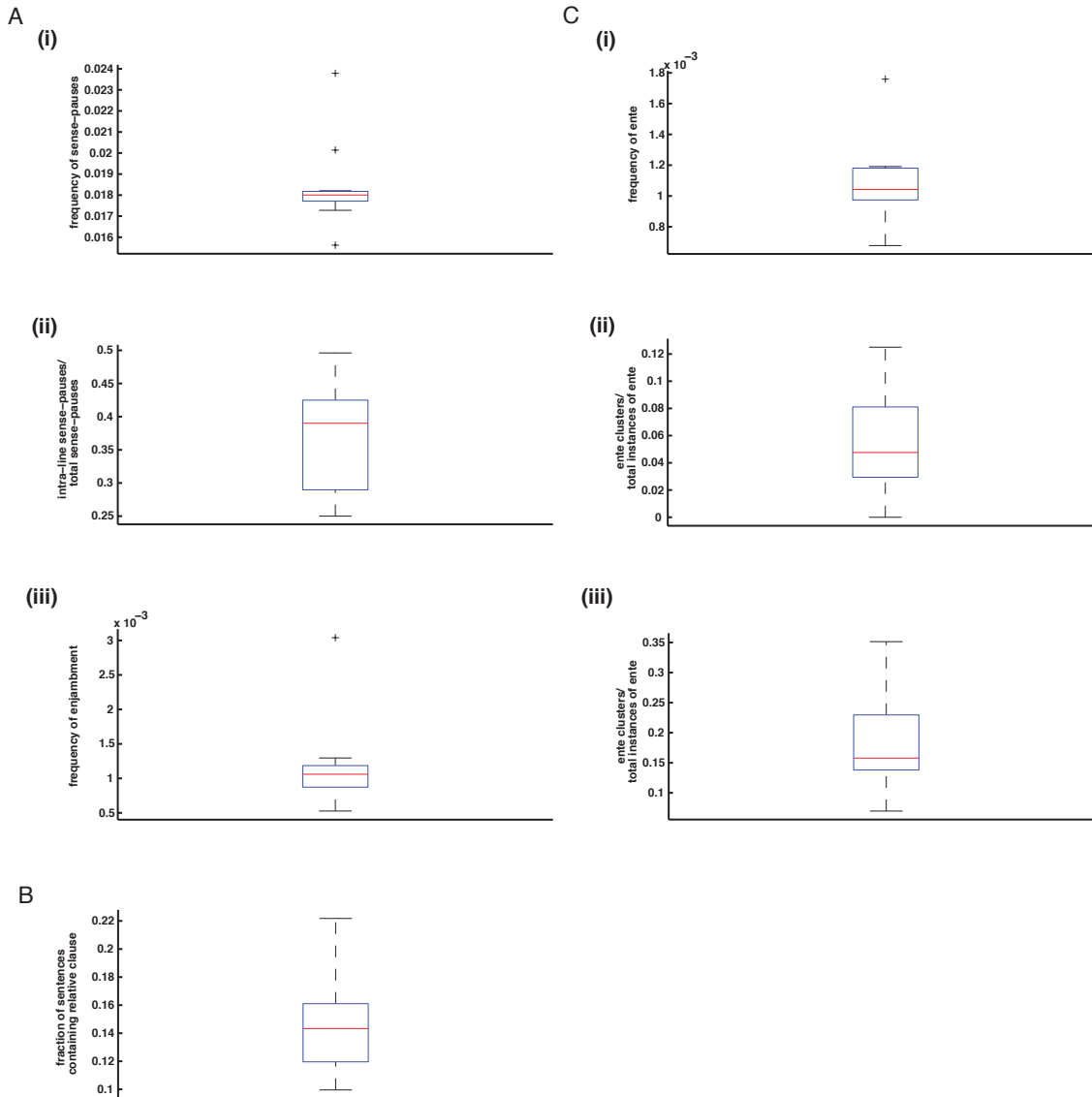


Fig. S1. Outliers in Senecan stylometric data. Box plots of the data presented in Fig. 2. (A, *i-iii*) correspond to Fig. 2A, *i-iii*, (B) corresponds to Fig. 2B, and (C, *i-iii*) correspond to Fig. 2D, *i* and *iv*. C, *ii* is for clusters within one line, C, *iii* for clusters within five lines. The red line denotes the median, the top and bottom of the blue box denote the 25th and 75th percentile, respectively, and the whiskers extend to the furthest non-outlier points. Outliers (black crosses) are defined as $> Q3 + 1.5IQR$ or $< Q1 - 1.5IQR$, where Q is the quartile and IQR is the interquartile range.

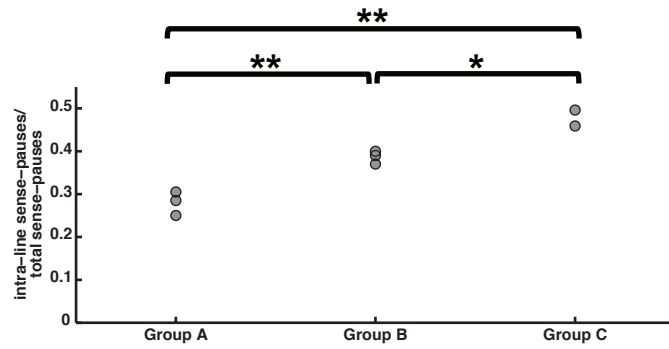


Fig. S2. Statistical analysis of Fitch's proposed groupings. Ratio of intra-line to total sense pauses for putatively early (group A), middle (group B), and late (group C) tragedies. Groupings follow Fitch 1981; sense-pauses were tabulated computationally using Peiper and Richter's text. At least one group is significantly different; $p < 0.001$ by a one-way ANOVA. Pairwise comparisons were made using a post-hoc Tukey HSD test; * $p < 0.05$, ** $p < 0.01$.

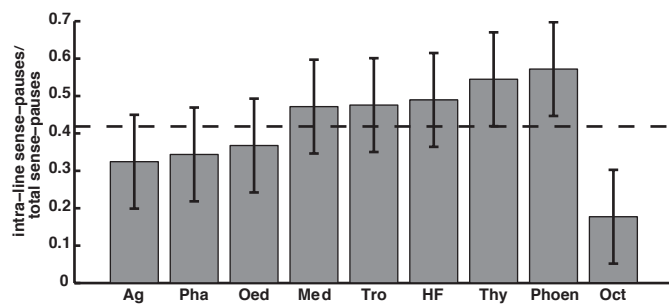


Fig. S3. Sense-pauses in Giardina's Seneca. Ratio of intra-line to total sense-pauses. Statistics for the eight authentic tragedies are reprinted from Fitch 1981. The ratio in the *Octavia* was determined by manual tabulation using Giardina's text. The dotted line denotes the mean; error bars denote one SD.

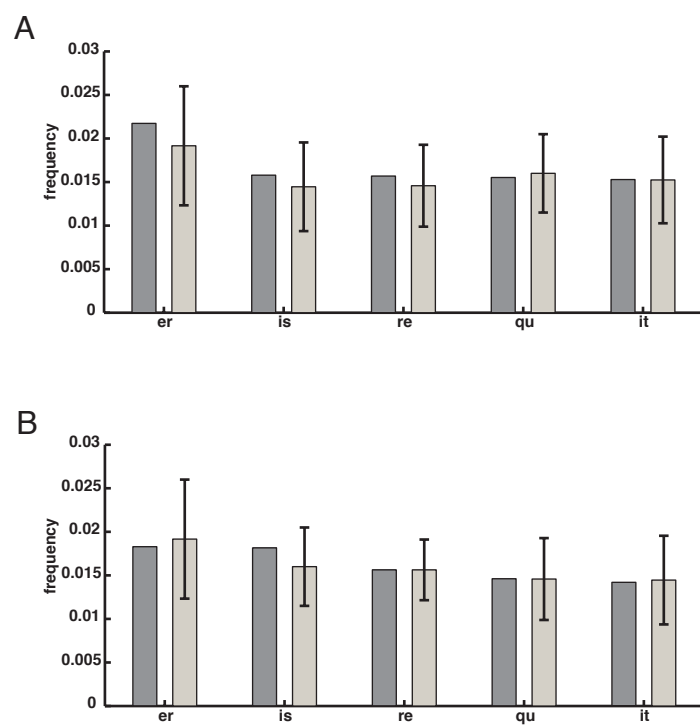


Fig. S4. Bigram analysis of the *Octavia* and *Hercules Oetaeus*. Per-character frequencies of the five most common bigrams in (A) *Octavia* and (B) *Hercules Oetaeus* (gray bars). Beige bars show the mean frequency of each n-gram across the 10 tragedies; error bars denote one SD.

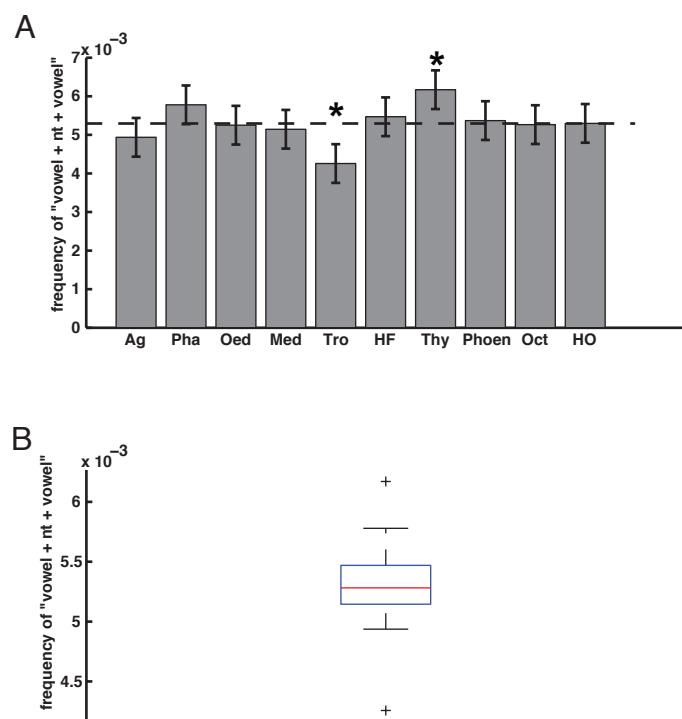


Fig. S5. Co-occurrences of "nt" with vowels. (A) Per-character frequency of the 20 combinations of the form "vowel + nt + vowel." Error bars indicate one SD across the 10 tragedies. (B) Box plot of the data in A.

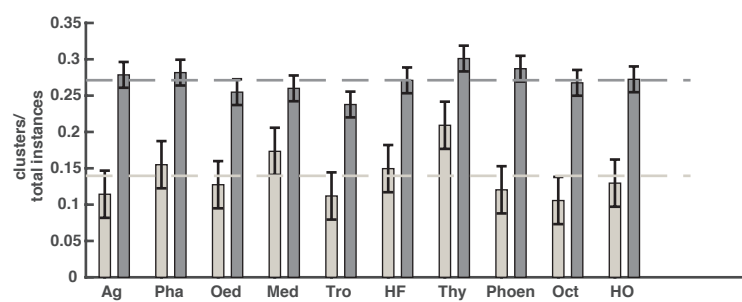


Fig. S6. Clusters of "vowel + nt + vowel" four-grams. Fraction of instances of "vowel + nt + vowel" four-grams that occur in clusters within each tragedy. The beige bars indicate instances within one line of each other, the gray bars within three.

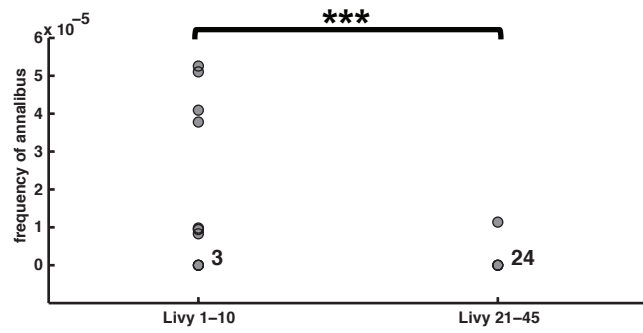


Fig. S7. Distribution of *annalibus* in Livy. Frequency of *annalibus* between the first decade (left) and subsequent (right) books of Livy. In multiple books of Livy the frequency of *annalibus* was 0 (indicated by the superscripts). *** $p < 0.001$ by a two-tailed unpaired t-test.

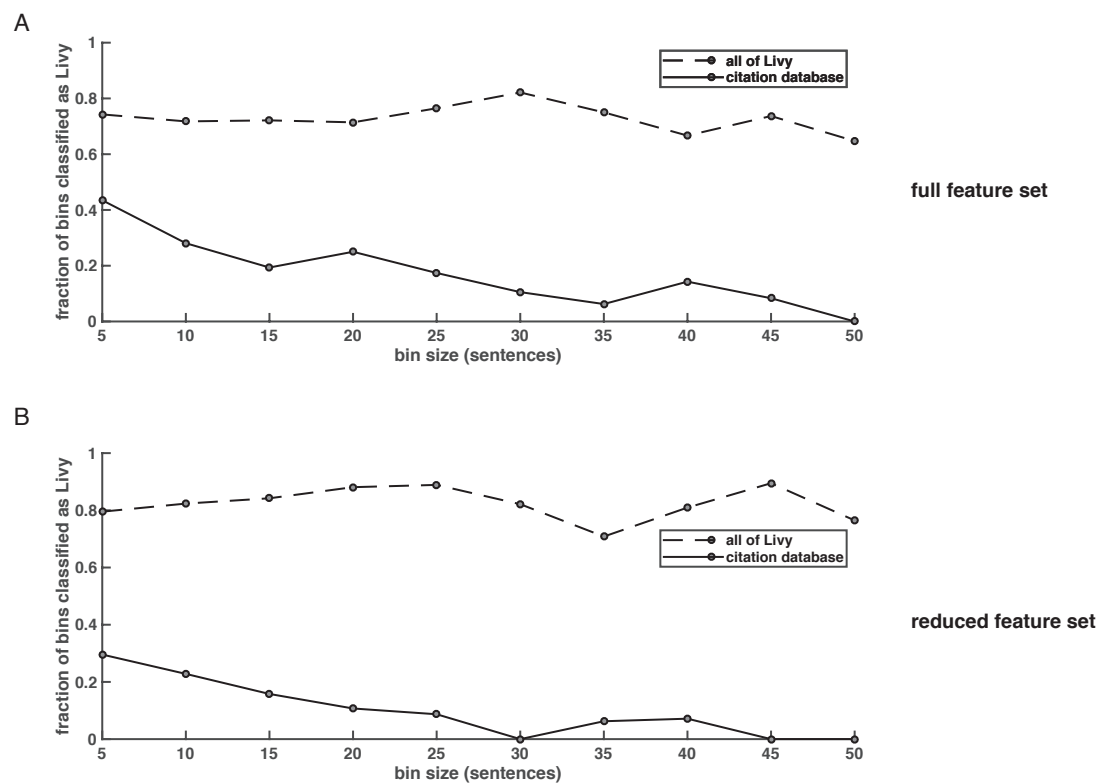


Fig. S8. Bin size and classifier performance. Fraction of bins from bulk Livy (dotted line) and the citation database (solid line) classified as Livian for bins of five sentences to 50 sentences using (A) the full set of 25 features and (B) the reduced set of five features.

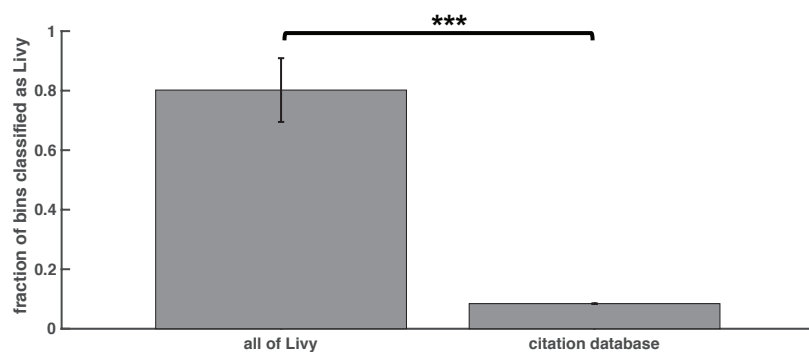


Fig. S9. Analysis of Livian citations using reduced feature set. Fraction of bins (random aggregates of 20 sentences) classified as Livian from bulk Livian material (left) and from the citation database (right) by a one-class SVM. Results are the mean \pm one SD of 35 leave-one-out cross-validation experiments. *** $p < 0.001$ by a two-tailed unpaired t-test.

A small set of stylometric features differentiates Latin prose and verse

Pramit Chaudhuri 

Department of Classics, University of Texas at Austin, TX, USA

Tathagata Dasgupta¹ and Joseph P. Dexter 

Department of Systems Biology, Harvard Medical School, MA, USA

Krithika Iyer²

Plano East Senior High School, TX, USA and Center for Excellence in Education, Research Science Institute, VA, US

Abstract

Identifying the stylistic signatures characteristic of different genres is of central importance to literary theory and criticism. In this article we report a large-scale computational analysis of Latin prose and verse using a combination of quantitative stylistics and supervised machine learning. We train a set of classifiers to differentiate prose and poetry with high accuracy (>97%) based on a set of twenty-six text-based, primarily syntactic features and rank the relative importance of these features to identify a low-dimensional set still sufficient to achieve excellent classifier performance. This analysis demonstrates that Latin prose and verse can be classified effectively using just three top features. From examination of the highly ranked features, we observe that measures of the hypotactic style favored in Latin prose (i.e. subordinating constructions in complex sentences, such as relative clauses) are especially useful for classification.

Correspondence:

Joseph P. Dexter, 200
Longwood Avenue, Boston,
MA 02115, USA.

E-mail:

jdexter@fas.harvard.edu

Interrogator: *In the first line of your sonnet which reads ‘Shall I compare thee to a summer’s day,’ would not ‘a spring day’ do as well or better?*

Witness: *It wouldn’t scan.*

Interrogator: *How about ‘a winter’s day,’ That would scan all right.*

Witness: *Yes, but nobody wants to be compared to a winter’s day.*

Interrogator: *Would you say Mr. Pickwick reminded you of Christmas?*

Witness: *In a way.*

Interrogator: *Yet Christmas is a winter’s day, and I do not think Mr. Pickwick would mind the comparison.*

Witness: *I don’t think you’re serious. By a winter’s day one means a typical winter’s day, rather than a special one like Christmas.*

A. M. Turing, ‘Computing Machinery and Intelligence’ (1950)

1 Introduction

The differences between prose and verse are often both numerous and straightforward. Meter, rhyme, form, tone, appearance on the page—any one of these features can be a decisive, instantaneous indicator of a text’s poetic quality. Nor is advanced

training in the humanities or creative writing typically required to tell poetry from prose; almost everyone has an intuitive appreciation that (for instance) rap lyrics and a political speech are quite different, over and above any differences in content. It is less easy, however, to explain precisely how poetry differs from prose, especially when standard formal features such as meter or rhyme are set aside, as in much free verse, or when prose writing favors rhetorical techniques typically associated more closely with poetry. For many readers, the distinction will come down to 'I know it when I see it,' rather than any ironclad criteria. Indeed, the question of what makes poetry poetic is one almost as old as literary theory itself, preoccupying critics from Horace to the Russian Formalists and many others besides. As the epigraph above attests, however, the question is of broader interest than to scholars of literature alone. Appearing in a well-known paper by Alan Turing, the passage—an imaginary dialogue between a computer ('Witness') and human ('Interrogator')—suggests that an understanding of the nature and function of poetry is paradigmatic for any convincing claim to artificial intelligence. For a machine to qualify as genuinely intelligent, it needs to do more than merely understand metrical rules ('it wouldn't scan'); rather, the computer would require the suppleness to grasp and appropriately respond to emotions, literary references, and other elements of meaning.

Against this background, it may come as less of a surprise that the rarefied realm of poetry, with its manifold hermeneutical challenges, has furnished a variety of problems of interest to contemporary computer scientists. In particular, the classification of prose and verse using machine learning has attracted attention as an initial avenue for integrating literary study and sophisticated computation. Although the basic task of distinguishing poetry and prose would seem to have a greater affinity with the rudimentary beginning of Turing's dialogue than its more ambitious end, classification promises more than an early step along the path toward artificial intelligence. As this article shows, the use of machine learning can also tease out certain subtle characteristics underlying prose or verse

and provide a quantitative profile of these different forms of expression.

A preliminary approach to prose/verse classification might focus on metrical features. The problem with such a line of attack, however, is the general absence of meter in prose texts, so that any classifier would trivially recapitulate manual annotations or the results of automated scansion programs.³ More promising strategies can be divided into two broad categories: image-based and text-based. Image-based approaches rely on the typically distinct appearance of poetry on the page; although they have proven useful for certain tasks such as data mining and document organization (Hanauer, 1996; Lorang *et al.*, 2015), image-based classifiers do not invite follow-up investigation of more intricate literary questions or integration with traditional modes of criticism. In contrast, text-based approaches are useful not only for binary classification but also for analysis (and can even enable novel modes of composition). Classification of prose and verse written in English has been accomplished using a range of machine learning algorithms and linguistic features (Hanauer, 1996; Tizhoosh and Dara, 2006; Tizhoosh *et al.*, 2008; Kumar and Minz, 2014). Successful analyses have also been reported for various premodern and non-Western literary traditions, including ecclesiastical Latin and Malay poetry (Manjavacas *et al.*, 2017; Jamal *et al.*, 2012). The documentation for the technical software package *Mathematica* even includes a workflow (with Shakespeare as a case study) for prose/verse classification without extensive user intervention or judgement.⁴ Particular attention has been devoted to poetic composition aided by machine learning on literary corpora, with notable recent examples including Swift-Speare (for generating pseudo-Shakespearean poetry) and DeepBeat (for rap lyrics) (Matias, 2010; Malmi *et al.*, 2016).

To the best of our knowledge, however, no similar analysis has been done for a classical literary tradition. In this article we report a large-scale characterization of Latin prose and verse using stylometric analysis and supervised machine learning. We train a set of classifiers that can differentiate prose and poetry very effectively (i.e. with accuracy values

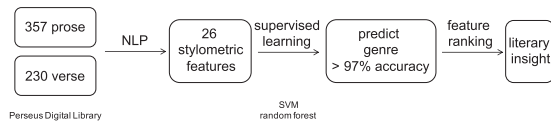


Fig. 1 Workflow for prose/verse classification. Twenty-six stylistometric features (Table 1) were calculated for 587 Latin text files (drawn from the Perseus Digital Library and further processed by the Tesseract Project) using custom heuristics. Two supervised learning algorithms (RF and linear SVM) were used to classify the text files as prose or verse, and statistical feature ranking was performed to gain insight into the stylistometric features that best distinguish the genres.

of >97%) using a set of twenty-six text-based features (Fig. 1). We then rank the relative statistical importance of the features to identify a low-dimensional set still sufficient for classification and offer detailed literary interpretations of the possible significance of those features. In particular, we find that measures of hypotactic style favored in much Latin prose (i.e. the use of subordinating constructions in complex sentences, such as relative clauses) often rank highly.

Our research leverages the pioneering work of the Perseus Digital Library, further facilitated by the Tesseract Project, to digitize almost all extant Greek and Latin literary texts (Crane, 1996; Coffee *et al.*, 2012). We obtained a set of 587 digitized Latin text files, which we divided into prose (357 files, ca. 112 works) and verse (230 files, ca. 94 works) following standard generic conventions. Each file typically contains either a whole work (e.g. an oratorical speech, such as Cicero’s *Pro Caelio*) or, in cases where a work is divided into individual books, one book (e.g. one of the three books of Caesar’s *De Bello Gallico* or one of the twelve of Vergil’s *Aeneid*). The set of texts (a full list of which is provided in the Appendix) includes almost all major surviving works of classical Latin literature, with the exception of substantially prosimetric works such as Boethius’ *Consolatio Philosophiae* or Seneca’s *Apocolocyntosis*, the latter of which is discussed below. (Petronius’ *Satyricon* was labeled as prose given that its prose content significantly exceeds its verse content.) The

chronological scope of the material is expansive, ranging from the comic plays of Plautus and fragments of Ennius' epic *Annales* (c. 200 BCE) to the poems of Ennodius (c. 500 CE) for verse, and from Cicero's early speeches (c. 80 BCE) to Jerome's letters (c. 400 CE) for prose. The generic scope is also broad. Represented in the corpus is epic (Vergil's *Aeneid*, Ovid's *Metamorphoses*) and didactic poetry (Lucretius' *De Rerum Natura*), tragedy (Seneca), comedy (Plautus, Terence), elegy (Propertius, Tibullus), historiography (Caesar, Livy, Tacitus, Suetonius), oratory (Cicero), philosophy (Cicero, Seneca), and technical writing (Vitruvius' *De Architectura*), to highlight only a handful of famous authors and works.

An important aspect of our approach is that it does not rely on syntactic parsing for feature extraction. Although development of a syntactic parser for Latin is an active area of research,⁵ natural language processing (NLP) research for Latin and other classical languages lags well behind efforts for English. We therefore devised a set of twenty-six features that could be computed without recourse to general-purpose syntactic parsing, for instance by tabulation of a signal word (e.g. a pronoun or conjunction) or signal *n*-gram (e.g. morphological endings and infixes indicating a particular grammatical function, such as the *-issim-* element characteristic of regular superlative adjectives and adverbs). Where a syntactical marker might have a homonym, we devised heuristics to disambiguate between them as far as possible. Certain features were deliberately made especially capacious or selective in response to the challenges posed by Latin morphology and complex syntax. In related research, we used a similar feature set to analyze and identify citations of fragmentary early historians in Livy's monumental history of Rome, a problem of major interest in Latin historiography (Dexter *et al.*, 2017). In total, the feature set is intended to provide a thorough (though inevitably incomplete) picture of Latin literary style, including many items that are of standard philological interest (e.g. usage of relative and other subordinate clauses) or that have proven useful for computational analysis of genre in other languages (e.g. prepositions) (Adams *et al.*, 2005; Jockers, 2013), and builds on

Burrows’ pioneering use of function words in literary stylometry (Burrows, 1987). Our approach was thus eclectic, derived from no single source and exploiting a range of features that have been applied to various languages and problems. In addition, we devised other features, including ones that were partial or noisy, to incorporate multiple types of evidence that could collectively capture diverse aspects of style. Table 1 lists the feature set divided into five broad grammatical categories: pronouns, non-content adjectives, conjunctions, subordinate clauses, and miscellaneous. We anticipate that our approach to feature extraction and literary machine learning will be applicable to other languages for which advanced NLP methods have not yet been developed.

A further goal of our research is to explore how stylometry and machine learning can support the practice of literary criticism. There is a long history of using stylometric analysis to address questions of authorship attribution and the dating of literary works in both classical and modern literary traditions (Mosteller and Wallace, 1964; Morton and Winspear, 1971; Marriott, 1979; Fitch, 1981; Holmes *et al.*, 2001; Vickers, 2004; Stamatatos, 2009; Forstall and Scheirer, 2010; Jockers and Witten, 2010; Stover *et al.*, 2016). Some recent work, however, has focused on the reapplication of stylometry, often involving machine learning methods, to address subtler literary critical questions. Notable examples include a statistical study of Shakespearean genre, integration of machine learning with traditional modes of criticism in the context of haiku, and an analysis of stylistic intertextuality in Latin tragedy and historiography (Hope and Witmore, 2010; Long and So, 2016; Dexter *et al.*, 2017), in addition to the development of pioneering frameworks of ‘distant reading’ (Moretti, 2013) and ‘macroanalysis’ (Jockers, 2013). Here we demonstrate that machine learning can decisively address a well-posed literary question, enabling us to identify large-scale stylistic signatures characteristic of Latin prose and verse genres. Moreover, we introduce methods for systematic ranking of feature importance, which are widely used in applications of machine learning outside

Table 1 Full set of Latin stylometric features

	Pronouns
1	Frequency of personal pronouns
2	Frequency of demonstrative pronouns
3	Frequency of <i>quidam</i>
4	Frequency of third-person reflexive pronouns
5	Frequency of <i>iste</i>
6	Frequency of <i>ipse</i>
7	Frequency of <i>idem</i>
	Non-content adjectives
8	Frequency of <i>alius</i>
	Conjunctions
9	Aggregate frequency of conjunctions
10	Frequency of <i>atque</i> followed by consonant
	Subordinate clauses
11	Frequency of conditional clauses
12	Frequency of <i>cum</i> clauses
13	Frequency of <i>quin</i> clauses
14	Frequency of <i>quominus</i> clauses
15	Frequency of <i>antequam</i> clauses
16	Frequency of <i>priusquam</i> clauses
17	Frequency of <i>dum</i> clauses
18	Fraction of sentences containing relative clause
19	Mean length of relative clauses
	Miscellaneous
20	Frequency of interrogative sentences
21	Frequency of selected vocatives
22	Frequency of superlatives
23	Frequency of <i>ut</i>
24	Frequency of selected gerunds and gerundives
25	Mean sentence length
26	Aggregate frequency of prepositions

Note: The twenty-six features are divided into five broad grammatical categories (pronouns, non-content adjectives, conjunctions, subordinate clauses, and miscellaneous).

of the digital humanities, to literary study (Chapelle and Vapnik, 1999; Guyon *et al.*, 2002; De la Torre and Vinyals, 2007; Grissa *et al.*, 2016).

2 Methods

2.1 Texts

All analyses were performed on a set of 587 Latin text files, most of which were originally digitized by the Perseus Digital Library (Crane, 1996) and further processed by the Tesserae Project. The set includes 357 prose files (ca. 112 works) and 230 verse files (ca. 94 works) and is publicly available at <https://github.com/tesserae/tesserae/tree/master/texts/la>, with the exception of six text files for the poet Phaedrus, which

were obtained directly from Perseus. The vast majority of the works are classical Latin. The full list of texts is provided in an Appendix at the end of the article (unless otherwise noted, all features were calculated for each individual book).

2.2 Computation of stylometric features

All NLP tasks were performed using JavaScript (ES2015). We computed a set of twenty-six Latin stylometric features for use in the prose/verse classifiers. All features are continuous, were computed without use of syntactic parsing, and fall into five broad categories (Table 1). The features in the first two categories (pronouns and non-content adjectives) were calculated by counting instances of the various inflected forms of the indicated Latin word(s). Tables of the inflected forms can be found in any standard textbook or reference grammar for Latin.⁶ A small number of feature calculations rely on modern editorial conventions, in particular punctuation, which has been exploited successfully in previous quantitative studies of classical literature (Clayman, 1981). In most cases the relevant punctuation is firm (e.g. a period or question mark) and is clearly implied by the syntax of the text, thereby reducing the likelihood of significant editorial differences, especially at scale.

Counts include either whole words or sequences of characters within words. For example, if counting instances of the polysemous word *ut*, which is both an adverb and a conjunction, we computed all appearances of the *n*-gram as a single word (e.g. *ut geniti*, *ut educati*, *ut cogniti essent*, not *Turnus rex Rutulorum*.) If counting (for instance) standard superlative forms, however, we computed all appearances of the relevant *n*-gram as a part of a word (*opulentissima*). Counts of whole words include both capitalized and lowercase forms, as well as instances with the enclitic *-que* (e.g. *utque* and *perque*), unless the enclitic produced another common Latin word (e.g. the number *quinque* instead of the conjunction *quin* with the enclitic *-que*) or the word was already included in the feature list (e.g. *atque* was not double-counted as *atque* and *at* + *que*). All frequencies are per character.

2.3 Conjunctions

- Conjunctions were computed by counting all instances of *ac*, *ast*, *at*, *atque*, *aut*, *autem*, *donec*, *dum*, *dummodo*, *enim*, *et*, *etenim*, *etiam*, *etiamtum*, *etiamtunc*, *nam*, *namque*, *nanque*, *neque*, *postquam*, *quamquam*, *quanquam*, *-que*, *quia*, *quocirca*, *sed*, *set*, *tamen*, *uel*, *uerumtamen*, *ueruntamen*, *utrumnam*, *vel*, *verumtamen*, and *veruntamen*.
- Frequency of *atque* followed by a consonant was computed by counting all instances of *atque* immediately followed by a word that begins with a consonant other than h (as h does not prevent elision).

2.4 Subordinate clauses

- Conditional clauses were computed by counting all instances of the conditional conjunctions *dummodo*, *nisi*, *quodsi*, *si*, and *sin*.
- *cum* clauses (where *cum* is an adverb or conjunction, but not a preposition) were computed by counting all instances of *cum* that are not immediately followed by a word ending in *-a*, *-e*, *-i*, *-o*, *-u*, *-is*, *-ibus*, *-ebus*, *-obus*, or *-ubus*. These restrictions were applied to exclude instances of *cum* as a preposition (which is followed by nouns in the ablative case, the inflected endings of which are listed above).
- *quin* clauses were computed by counting all instances of *quin*.
- *quominus* clauses were computed by counting all instances of *quominus* and *quo minus*.
- *antequam* clauses were computed by counting all instances of *antequam* and *ante quam*.
- *priusquam* clauses were computed by counting all instances of *priusquam* and *prius quam*.
- *dum* clauses were computed by counting all instances of *dum*.
- The fraction of non-interrogative sentences containing at least one relative clause was computed by identifying sentences that are both non-interrogative (i.e. ending with a punctuation mark other than '?') and have at least one form of the Latin relative pronoun (*qui*, *cuius*, *cui*, *quem*, *quo*, *quae*, *quam*, *qua*, *quod*, *quorum*,

quibus, quos, quarum, or quas). Interrogative sentences were excluded to obviate the need for semantic parsing of relative and interrogative pronouns, which are often morphologically identical.

- The average length of relative clauses was computed by counting the number of characters, excluding spaces and punctuation, in relative clauses, identified as beginning with a relative pronoun and ending at the next punctuation mark.

2.5 Miscellaneous

- (Direct) interrogative sentences were computed by counting all instances of a sentence ending in a question mark.
- Vocatives were computed by counting all instances of 'o' followed by a single word ending in *-a*, *-e*, *-i*, *-u*, *-ae*, *-es*, *-um*, or *-us*. The limitations were applied to exclude stand-alone instances of 'o' (an exclamation of surprise as well as part of a direct address), but to include instances of 'o' followed by a word with a stand-alone vocative case ending.
- Regular superlative adjectives and adverbs were computed by counting all instances of *-issim-* within a word. The method excludes certain common superlatives such as *maximus* or *optimus*, which would be difficult to capture precisely without also incorporating proper names (e.g. Fabius Maximus or Jupiter Optimus Maximus).
- Frequency of *ut* (where *ut* is an adverb or a conjunction) was computed by counting all instances of *ut*.
- The limited subset of gerunds and gerundives was computed by counting all instances of *-ndus*, *-ndum*, *-ndarum*, and *-ndorum*. The restriction was designed to exclude the many verb forms that share the same letter sequence as the characteristic gerundival ending (e.g. *defendo* and *pendo*), though at the cost of also excluding the majority of the inflected forms of the gerund and gerundive. The common adverb *nondum* was excluded from this count.

Erroneous inclusion of words such as the adjective *blandus* or noun *mundus* was assumed not to vitiate the count.

- The average length of sentences was computed by counting the number of characters, excluding spaces and punctuation, in sentences ending in a ‘.’, ‘?’, or ‘!’. We excluded any periods occurring after a single stand-alone character, since such instances are typically an abbreviation of a proper name rather than a sentence-end, and periods occurring after other common abbreviations (e.g. Aug., Cn., Kal., or common multi-letter Roman numerals followed by a period).
- Prepositions were computed by counting all instances of *ab*, *abs*, *absque*, *apud*, *cis*, *de*, *e*, *erga*, *ex*, *inter*, *ob*, *penes*, *per*, *praeter*, *pro*, *propter*, *sub*, *tenuis*, and *trans*. Prepositions that may function as adverbs were excluded.

2.6 Error analysis

As described above, certain features could not be computed exactly and instead were estimated using various heuristics. To assess the effectiveness of these heuristics, we performed a manual error analysis of three features: frequency of regular superlatives, frequency of *cum* clauses, and frequency of the enclitic conjunction *-que*, which was a subset of our aggregated conjunctions feature. We analyzed the features in a small set of Latin prose and verse texts: Seneca’s *Phoenissae* and the *Octavia* (verse) and Livy 22.1-15 (prose) for relative clauses, *Aeneid* 1 (verse) and Livy 22.1-15 for superlatives and *-que*, and Livy 22.1-15 for *cum* clauses. Table 2 lists the precision and recall of each heuristic in the texts analyzed.

2.7 Supervised machine learning

Prior to classification all features were rescaled to have minimum value 0 and maximum value 1. All supervised learning tasks were performed using Python 2.7. We used the scikit-learn implementations of the binary support vector machine (SVM) and random forest (RF) classifiers (Pedregosa *et al.*, 2011). For the linear SVM, we set $C=0.5$, which is the default value in the scikit-learn package. For the RF classifier, feature ranking was according to Gini importance (Breiman and Cutler, 2008); ranking for

Table 2 Error analysis of selected features

Feature	Text	TP	FP	FN	Precision	Recall
Relatives	<i>Phoenissae</i> (V)	73	9	19	0.89	0.79
Relatives	<i>Octavia</i> (V)	103	7	7	0.94	0.94
Relatives	Livy 22.1-15 (P)	90	28	9	0.76	0.91
Superlatives	<i>Aeneid</i> 1 (V)	5	1	0	0.83	1
Superlatives	Livy 22.1-15 (P)	5	0	0	1	1
<i>cum</i> clauses	Livy 22.1-15 (P)	14	0	13	1	0.52
<i>-que</i>	<i>Aeneid</i> 1 (V)	280	19	0	0.94	1
<i>-que</i>	Livy 22.1-15 (P)	154	41	0	0.79	1

Notes: Table 2 summarizes the performance of four heuristics used for feature extraction on a sample of Latin texts. P and V indicate prose and verse, respectively, and TP, FP, and FN refer to true positives, false positives, and false negatives, respectively. Data on relative pronouns are derived from the feature ‘mean length of relative clause’ (i.e. the error being analyzed is the incorrect identification of the relative pronoun using the heuristics underlying the feature). *-que* is not a stand-alone feature but is the only item in the aggregate frequency of conjunctions that could not be calculated exactly.

the linear SVM was determined by inspection of the set of weights of each support vector associated with the features' contributions (Chang and Lin, 2008). Unless otherwise noted, all results are based on five-fold cross-validation.

2.8 Code availability

All code is freely and publicly available at <https://github.com/qcrit>.

results demonstrate that Latin prose and verse can be differentiated using stylometric features and prompted us to examine the relative importance of individual features, as discussed below. In addition, we confirmed that the RF model could classify texts artificially partitioned into 500-word chunks as prose or verse (mean cross-fold accuracy >90%), indicating that classifier performance is not strongly influenced by text length.⁷

3 Results

We report a set of supervised learning classifiers that can distinguish Latin prose and verse with high accuracy and a literary critical examination of the stylistic features most useful for prose/verse classification. Table 1 lists the full set of twenty-six stylistic features, and an outline of the computational workflow is shown in Fig. 1.

3.1 Classification of prose and verse using the full set of stylometric features

We first attempted classification of Latin prose and poetry using all twenty-six features and two classification algorithms (RF and SVM with a linear kernel). Table 3 summarizes the results with five-fold cross-validation. With RF a total of fourteen texts were misclassified, and the mean accuracy across the folds was 97.6%; with the linear SVM, a total of thirteen texts were misclassified, and the mean accuracy across the folds was 97.8%. These

3.2 Feature ranking identifies a small set of stylometric features sufficient for high-accuracy classification

We ranked the twenty-six features according to their contribution to classifier performance using both RF and linear SVM; Table 4 lists the top ten features by Gini importance (RF) or by the absolute value of the feature weight (linear SVM). Six of the top ten features are common between the models, as are three of the top four (fraction of sentences containing a relative clause, frequency of regular superlatives, and frequency of *quidam*), suggesting that our analysis identified a reproducible set of critical features. Furthermore, the three and five highest-ranked features alone are sufficient for prose/verse classification with >95% accuracy (Table 5). The literary and linguistic significance of the top features is reviewed in detail in the Discussion.

3.3 Classification of the Apocolocytosis

The *Apocolocyntosis* ('Pumpkinification') is a satire on the deification of the emperor Claudius. Traditionally

Table 3 Performance of prose/verse classification using full feature set with five-fold cross-validation

	Fold 1 (%)	Fold 2 (%)	Fold 3 (%)	Fold 4 (%)	Fold 5 (%)	Mean (%)	SD (%)	Accuracy (%)	F1 score (%)
RF	97.5	98.3	96.6	99.1	96.6	97.6	1.1	97.6	97.5
SVM	98.3	100	97.4	96.6	96.6	97.8	1.4	97.8	97.7

Note: Table 3 lists the accuracy for each fold, along with the mean and SD across folds, overall accuracy, and overall $F1$ score (macro-averaged).

Table 4 List of highly ranked features

Ranking	RF	Gini importance	SVM (linear kernel)	$ w $
1	Frequency of superlatives	0.223	Frequency of superlatives	2.69
2	Fraction of sentences containing relative clause	0.184	Frequency of <i>antequam</i> clauses	1.65
3	Frequency of <i>quidam</i>	0.144	Fraction of sentences containing relative clause	1.56
4	Frequency of <i>idem</i>	0.0922	Frequency of <i>quidam</i>	1.53
5	Aggregate frequency of prepositions	0.0678	Frequency of <i>alius</i>	1.50
6	Frequency of selected gerunds and gerundives	0.0634	Frequency of <i>idem</i>	1.39
7	Frequency of <i>dum</i> clauses	0.0334	Frequency of personal pronouns	1.28
8	Frequency of selected vocatives	0.0331	Frequency of <i>iste</i>	1.22
9	Frequency of <i>ut</i>	0.0312	Frequency of <i>dum</i> clauses	1.07
10	Frequency of <i>cum</i> clauses	0.0257	Aggregate frequency of prepositions	1.06

Note: For RF the features are ranked by Gini importance; for the linear SVM, they are ranked by the absolute value of the weight of the support vector associated with the contribution of the feature.

Table 5 Accuracy of prose/verse classification using RF with reduced feature sets

	Fold 1 (%)	Fold 2 (%)	Fold 3 (%)	Fold 4 (%)	Fold 5 (%)	Mean (%)	SD (%)	Accuracy (%)	F1 score (%)
All	97.5	98.3	96.6	99.1	96.6	97.6	1.1	97.6	97.5
Top 10	95.8	97.5	96.6	96.6	95.7	96.4	0.7	96.4	96.2
Top 5	98.3	97.5	95.7	97.4	94.0	96.6	1.7	96.8	96.6
Top 3	96.6	94.1	94.9	96.6	94.0	95.2	1.3	95.2	95.0

Note: For the three reduced feature sets, Table 5 lists the accuracy for each fold, along with the mean and SD across folds, overall accuracy, and overall *F1* score (macro-averaged). The top features (by Gini importance with RF classification) are given in Table 4.

attributed to the statesman and philosopher Seneca, its date of composition is uncertain but likely to be soon after Claudius' death in 54 CE. The work is written in both prose and verse: the prose sections narrate the emperor's journey and encounters in the afterlife and are interspersed with passages of poetry in a high linguistic register. The combination makes the text an especially attractive candidate for testing automated differentiation of prose and verse. The various sections of the text (again drawn from the Perseus Digital Library) were aggregated into two bins, one of prose and the other of verse, for classification using the full set of twenty-six features. Both bins were correctly classified using an RF classifier trained on the full set of 587 texts.

4 Discussion

Our stylistometric features collectively capture different aspects of the prosaic or poetic quality of a text, which accounts for their effectiveness as a combined set. Beyond this collective effectiveness, certain specific features stand out as having an intuitively greater suitability for one form of expression over the other. Two of the top five features by both rankings, for instance, point toward the relatively fuller and less restrictive scope of prose sentences in comparison to verse. Although not all Latin prose fits that characterization—prose writings, especially in the first-century AD, sometimes exploited the short, pointed sentence, and certain genres in any period,

Table 6 List of texts misclassified by both RF and linear SVM

Author	Text	Book
Lucretius	<i>De rerum natura</i>	1
Lucretius	<i>De rerum natura</i>	2
Lucretius	<i>De rerum natura</i>	3
Lucretius	<i>De rerum natura</i>	4
Lucretius	<i>De rerum natura</i>	5
Lucretius	<i>De rerum natura</i>	6
Manilius	<i>Astronomica</i>	1
Manilius	<i>Astronomica</i>	2
Sallust	<i>Historiae</i>	All
Tacitus	<i>Annales</i>	12

Note: Ten texts (eight verse and two prose) were misclassified by both models.

such as legal writings and literary commentary, could be fairly compact in expression (Kennedy, 1994; Adams *et al.*, 2005)—the general trend is nevertheless sufficient for texts to be classified effectively using only the highest-ranked features.

One of the most important features among the twenty-six for distinguishing between prose and verse is the fraction of non-interrogative sentences containing at least one relative clause (ranked second using RF, third using SVM). This finding reflects a common difference in the syntactical structures of the two forms. Prose tolerates longer sentences, which can be broken down into smaller clauses using a variety of subordinating constructions. The most common such construction is the relative clause, which hinges on the relative pronoun, 'who' or 'which' (in Latin the various inflected forms of *qui*). Relative clauses—like other subordinating constructions, such as conditional clauses, purpose clauses ('in order to'), and many further types—organize a thought into main and subsidiary elements, prior and latter actions, actual and contingent events, etc. With little restriction on sentence length, prose authors could regularly avail themselves of complex subordination to qualify a thought, or, more simply, to avoid the monotony of a series of parallel clauses. It is prose authors' customary favoring of this hypotactic, as opposed to paratactic, style that partly accounts for the prominence of the relative clause feature. This is not to say that verse is paratactic—far from it—

but rather that the shorter sentences more often used by poets (mean length 102.3 ± 38.1 characters for the verse files, compared to 128.8 ± 37.1 for prose) reduced the need for extensive subordination, and hence relative clauses. Furthermore, one specific use of the relative pronoun may favor prose over verse: *qui* (or its inflections) often appears at the beginning of a sentence where it refers to an antecedent in the previous sentence, a usage known as the connecting relative. Although common to both prose and poetry, it is a frequent feature of Caesarian and other prose texts (Mayer, 2005; Spevak, 2010). Additional corroboration of the importance of hypotactic markers can be found among the top ten features: frequency of *ut* (ranked ninth using RF) and of *antequam* clauses (ranked second using SVM). In the former case, the word *ut* does have some non-subordinating uses, but a very large number of occurrences introduce one of a range of dependent clauses indicating purpose, result, or command, among other types. In the latter case, the adverb *antequam* ('sooner than', 'before') introduces certain temporal clauses. The feature rankings thus point to three potential markers of hypotaxis (*qui*, *ut*, and *antequam*) as playing an important role in the differentiation of prose and verse.

A different kind of expansiveness (or, conversely, compactness) is likely to account for the top ranking of superlative adjectives and adverbs, which again are enriched in prose texts. In this case, the relevant unit of analysis is not the sentence but rather the verse line. Latin poetry was metrical or quantitative and was structured according to patterns of syllable weight, primarily, and stress, secondarily. These patterns allowed for a certain number of syllables in the verse line, with twelve to seventeen syllables, for example, allowed in epic. At a mechanical level, then, verse composition partially consisted of the art of fitting choice words into interesting configurations while conforming to the metrical rules of the particular genre or subgenre (epic, tragedy, elegy, choral ode, etc.). The most common form of the superlative adjective in Latin adds the combined infix and ending *-issimus* (or its various inflections) to the stem of the base adjective, so that the resulting word is at a minimum four syllables long

(e.g. *fortissimus*, 'very brave') and in many cases five or more. Such a word can be accommodated into most verse meters, but it occupies a significant proportion of the line. Some words are especially long and would take up almost half of an entire verse line. The heptasyllabic word *tumultuosissimus* ('very tumultuous'), for instance, can be used by the historian Livy (2.10) without repercussion, but it presents a serious challenge to any poet, over and above the fact that its syllabic pattern prevents the word from being used at all in certain meters. In contrast to the subordinating features above, the frequency of superlative adjectives and adverbs is thus likely to be affected by meter rather than syntax. Both types of feature, however, are more characteristic of prose. Mostly unrestricted by metrical rules and often highly elastic and hypotactic in syntax, prose could accommodate several of the top-ranked features with somewhat greater ease than poetry.

Of the 587 texts in the corpus, a mere ten were misclassified (eight verse, two prose) by both models using the full feature set (Table 6). These shared misclassifications merit some explanation. The largest and most interesting category of misclassified texts contains didactic poetry, in particular the philosophical poems of Lucretius and Manilius. Both works are naturally influenced by philosophical prose, which plausibly explains features such as their longer-than-average sentence length compared to other hexameter works. The effects of that influence may partially account for the misclassifications. Consistent with that hypothesis, Vergil's *Georgics*, a didactic poem on an agricultural rather than conventionally philosophical theme, was correctly classified. In contrast, there is no clear reason for the misclassification of Sallust's *Historiae* and Book 12 of Tacitus' *Annales*; none of the other extant books of the *Annales*, and no other work by either author, was misclassified. In such mysterious cases it may be unprofitable to speculate on an explanation: the addition of a new feature or even the enhancement of an existing feature might be all that is required for a correct classification.

Insofar as the core objective of the experiment was to distinguish between Latin prose and verse, the achievement of >97% accuracy is a notable

success. Since Latin poetry comprises a highly structured set of generic norms defined by meter, thematic content, and other features, it may seem to be a more straightforward candidate for classification than, say, much English poetry, especially those types using looser or non-existent meter and prosaic vocabulary. As discussed earlier, however, for all its structure Latin verse is also characterized by a remarkable diversity of metrical forms and subject matter, ranging from tightly constrained lyric to highly flexible drama. Moreover, poetry and prose often overlap in content: epic and historiography share long narratives about kings and battles, while letters are written in both prose and elegiac verse. Despite those challenges, our approach shows that Latin poetry—with extensive formal diversity and topical breadth—is nevertheless highly amenable to computational differentiation from prose. Also notable is the basis of this differentiation: mostly syntactic features, which were calculated without the aid of syntactic parsing. Attention to lexical markers as a proxy for syntactical constructions, coupled with heuristics to disambiguate between words of similar form, may offer a productive path forward for researchers working on literary traditions that lack the technical resources of English and other commercially important modern languages.

Furthermore, the introduction of feature ranking enabled us to identify the most salient features for the differentiation of prose and verse. Although routinely employed in other applications of machine learning, such as bioinformatics, it is an underutilized component of the toolkit of the digital humanist. Leveraging high-dimensional calculations well beyond the capacity of a human researcher, feature ranking can bring to light subtle yet important relationships between individual features and the data set as a whole. We hope that our application of feature ranking can provide a useful model for digital humanists seeking to extend the power of classification as a critical method.

Acknowledgements

The authors thank Jeffrey Flynt, Thomas Bolt, and Elizabeth Adams for assistance with development of

the Latin feature set. This work was conducted under the auspices of the Quantitative Criticism Lab (www.qcrit.org), an interdisciplinary project co-directed by J.P.D. and P.C. and supported by seed funding from the Office of the Provost at Dartmouth College, a Neukom Institute for Computational Science CompX Faculty Grant, and a National Endowment for the Humanities Digital Humanities Start-Up Grant (grant number HD-248410-16). J.P.D. was supported by a National Science Foundation Graduate Research Fellowship (grant number DGE1144152) and a Neukom Fellowship, and P.C. was supported by an American Council of Learned Societies Digital Innovation Fellowship and a New Directions Fellowship from the Andrew W. Mellon Foundation. K.I. was a Research Science Institute Scholar during the summer of 2016.

Appendix

The full list of texts is as follows (unless otherwise noted, all features were calculated for each individual book):

Verse texts: Anonymous, *Laudes Domini*; Catullus, *Carmina* (divided into three files: epic, elegy, and miscellaneous); Claudian, *Carmina Minora*, *De Bello Gildonico*, *De Bello Gothico*, *De Consulatu Stilichonis*, *De Raptu Proserpinae*, *Epithalamium De Nuptiis Honorii Augusti*, *Panegyricus Dictus Probino et Olybrio Consulibus*, *In Eutropium*, *In Rufinum*, *Panegyricus De Tertio Consulatu Honorii Augusti*, *Panegyricus De Quarto Consulatu Honorii Augusti*, *Panegyricus De Sexto Consulatu Honorii Augusti*, and *Panegyricus Dictus Manlio Theodoro Consuli*; Dracontius, *Romulea* 10; Ennius, *Annales*; Ennodius, *Carmina* (Books 1 and 2 combined); Horace, *Ars Poetica*, *Carmen Saeculare*, *Epistles* (Books 1 and 2 combined), *Epodes*, *Odes*, and *Satires*; Italicus, *Ilias Latina*; Juvenal, *Satires*; Juvenius, *Historia Evangelica*; Lucan, *Bellum Civile*; Lucretius, *De Rerum Natura*; Manilius, *Astronomica*; Martial, *Epigrams*; Ovid, *Amores*, *Ars Amatoria*, *Epistulae Ex Ponto*, *Fasti*, *Heroides*, *Ibis*, *Medicamina Faciei Femineae*, *Metamorphoses*, *Remedia Amoris*, and *Tristia*; Persius, *Satires*; Phaedrus, *Fabulae*;

Plautus, *Amphitruo*, *Asinaria*, *Aulularia*, *Bacchides*, *Captivi*, *Casina*, *Cistellaria*, *Curculio*, *Epidicus*, *Menaechmi*, *Mercator*, *Miles Gloriosus*, *Mostellaria*, *Persa*, *Poenulus*, *Pseudolus*, *Rudens*, *Stichus*, *Trinummus*, and *Truculentus*; Propertius, *Elegies* (Books 1–4 combined); Prudentius, *Apotheosis*, *Contra Symmachum*, *Dittochaeon*, *Epilogus*, *Hamartigenia*, and *Psychomachia*; Rutilius Namatianus, *De Reditu Suo*; Seneca, *Agamemnon*, *Hercules Furens*, *Hercules Oetaeus*, *Medea*, *Octavia*, *Oedipus*, *Phaedra*, *Phoenissae*, *Thyestes*, and *Troades*; Silius Italicus, *Punica*; Statius, *Achilleid* (Books 1 and 2 combined), *Silvae* and *Thebaid*; Terence, *Adelphi*, *Andria*, *Eunuchus*, *Heautontimorumenos*, *Hecyra*, and *Phormio*; Tibullus, *Elegies* (Books 1–3 combined); Valerius Flaccus, *Argonautica*; Vergil, *Aeneid*, *Eclogues*, and *Georgics*.

Prose texts: Ammianus, *Rerum Gestarum*; Apuleius, *Apologia*, *Florida*, and *Metamorphoses*; Augustine, *Epistulae* (Books 1–10, 11–20, 21–30, 31–40, 41–50, and 51–62); Julius Caesar, *De Bello Civili* and *De Bello Gallico*; Augustus Caesar, *Res Gestae Divi Augusti*; Celsus, *De Medicina*; Cicero, *Academica*, *Brutus*, *Cum Populo Gratias Egit*, *De Amicitia*, *De Divinatione*, *De Domo Sua*, *De Fato*, *De Finibus Bonorum et Malorum*, *De Haruspicum Responso*, *De Imperio Cn. Pompei*, *De Inventione*, *De Lege Agraria Contra Rullum*, *De Natura Deorum*, *De Officiis*, *De Optimo Genere Oratorum*, *De Oratore*, *De Partitione Oratoria*, *De Provinciis Consularibus*, *De Republica*, *De Senectute*, *Divinatio in C. Verrem* (Books 1, 2.1, 2.2, 2.3, 2.4, and 2.5), *Divinatio in Q. Caecilius*, *Epistulae ad Familiares*, *In Catilinam* (Books 1–4 combined), *In L. Pisonem*, *In Vatinius*, *Epistulae ad Atticum*, *Epistulae ad Brutum*, *Epistulae ad Quintum Fratrem*, *Lucullus*, *Orator*, *Paradoxa Stoicorum ad M. Brutum*, *Philippicae*, *Post Reditum in Senatu*, *Pro A. Caecina*, *Pro A. Cluentio*, *Pro Archia*, *Pro Balbo*, *Pro C. Rabirio*, *Pro C. Rabirio Postumo*, *Pro Fonteio*, *Pro L. Flacco*, *Pro Ligario*, *Pro M. Caelio*, *Pro Marcello*, *Pro Milone*, *Pro Murena*, *Pro Plancio*, *Pro Publio Quinctio*, *Pro Q. Roscio Comoedo*, *Pro Rege Deiotaro*, *Pro S. Roscio*, *Pro Scauro*, *Pro Sestio*, *Pro Sulla*, *Pro Tullio*, *Timaeus*, *Topica*, and *Tusculanae Disputationes*; Q. Cicero, *Commentariolum Petitionis*; Columella, *De Re*

- Rustica* (Books 1–9 only); Curtius Rufus, *Historiae Alexandri Magni*; Florus, *Epitomae De Tito Livio Bellorum Omnium Annorum DCC*; Gellius, *Noctes Atticae*; Jerome, *Epistulae* selections (Letters 1, 7, 14, 22, 38, 40, 43, 44, 45, 52, 54, 60, 77, 107, 117, 125, 127, and 128 combined); Livy, *Ab Urbe Condita* (Books 1–10, 21–30, 31–40, and 41–45); Marcus Minucius Felix, *Octavius*; Nepos, *Vitae*; Petronius, *Satyricon*; Pliny the Elder, *Naturalis Historia* (Books 1–5, 6–10, 11–15, 16–20, 21–25, 26–30, and 31–37); Pliny the Younger, *Epistulae*; Pseudo-Cicero, *In Sallustium*; Pseudo-Quintilian, *Declamationes Maiores*; Quintilian, *Institutio Oratoria*; Sallust, *Catilina*, *Historiae*, and *Jugurtha*; Scriptores Historiae Augustae, *Historia Augusta* (Books 1–5, 6–10, 11–15, and 16–21); Seneca the Younger, *Epistulae ad Lucilium* (Letters 1–10, 11–20, 21–30, 31–40, 41–50, 51–60, 61–70, 71–80, 81–90, 91–100, 101–110, 111–120, and 121–124), *De Beneficiis*, *De Brevitate Vitae*, *De Clementia*, *De Consolatione ad Helviam*, *De Consolatione ad Marciam*, *De Consolatione ad Polybium*, *De Constantia*, *De Ira*, *De Otio*, *De Providentia*, *De Tranquillitate Animi*, and *De Vita Beata*; Seneca the Elder, *Controversiae* (Books 1, 2, 7, 8, and 10), *Excerpta Controversiae*, *Fragmenta*, and *Suasoriae*; Suetonius, *De Vita Caesarum*; Tacitus, *Agricola*, *Annales*, *De Origine et Situ Germanorum*, *Dialogus de Oratoribus*, and *Historiae*; Tertullian, *Apologeticum* and *De Spectaculis*; Valerius Maximus, *Facta et Dicta Memorabilia*; and Vitruvius, *De Architectura*.
- ## References
- Adams, J. N., Lapidge, M., and Reinhardt, T. (2005). Introduction. In Reinhardt, T., Lapidge, M., and Adams, J. N. (eds), *Aspects of the Language of Latin Prose. Proceedings of the British Academy*, vol. 129. Oxford: Oxford University Press, pp. 1–36.
- Breiman L. and Cutler A. (2008). Random forests—classification manual. <http://www.math.usu.edu/~adele/forests/> (accessed 30 August 2017).
- Burrows, J. F. (1987). *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press.
- Chang, Y.-W. and Lin, C.-J. (2008). Feature ranking using linear SVM. In *JMLR: Workshop and Conference Proceedings*, vol. 3, pp. 53–64. <http://proceedings.mlr.press/v3/chang08a/chang08a.pdf> (accessed 30 August 2017).
- Chapelle, O. and Vapnik, V. (1999). Model selection for support vector machines. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, Denver, vol. 12, pp. 230–6.
- Clayman, D. L. (1981). Sentence length in Greek hexameter poetry. In Grotjahn, R. (ed.), *Hexameter Studies. Quantitative Linguistics 11*. Bochum: Brockmeyer, pp. 107–36.
- Coffee, N., Koenig, J.-P., Poornima, S., Ossewaarde, R., Forstall, C., and Jacobson, S. (2012). Intertextuality in the digital age. *Transactions of the American Philological Association*, 142(2): 383–422.
- Crane, G. (1996). Building a digital library: the Perseus Project as a case study in the humanities. In *Proceedings of the First ACM International Conference on Digital Libraries*, Bethesda, vol. 1, pp. 3–10.
- De la Torre, F. and Vinyals, O. (2007). Learning kernel expansions for image classification. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis. <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4270176> (accessed 30 August 2017).
- Dexter, J. P., Katz, T., Tripuraneni, N., Dasgupta, T., Kannan, A., Brofos, J. A., Bonilla Lopez, J. A., Schroeder, L. A., Casarez, A., Rabinovich, M., Haimson Lushkov, A., and Chaudhuri, P. (2017). Quantitative criticism of literary relationships. *Proceedings of the National Academy of Sciences United States of America*, 114(16): E3195–204.
- Fitch, J. (1981). Sense-pauses and relative dating in Seneca, Sophocles and Shakespeare. *American Journal of Philology*, 102(3): 289–307.
- Forstall, C. W. and Scheirer, W. J. (2010). Features from frequency: authorship and stylistic analysis using repetitive sound. In *Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science*, 1.2. Chicago, IL: University of Chicago.
- Grissa, D., Pétera, M., Brandolini, M., Amedeo, N., Comte, B., and Pujos-Guillot, E. (2016). Feature selection methods for early predictive biomarker discovery using untargeted metabolomic data. *Frontiers in Molecular Biosciences*, 3: 30.
- Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1–3): 389–422.

- Hanauer, D. (1996). Integration of phonetic and graphic features in poetic text categorization judgements. *Poetics*, 23(5): 363–80.
- Holmes D. I., Robertson, M., and Paez, R. (2001). Stephen Crane and the *New-York Tribune*: a case study in traditional and non-traditional authorship attribution. *Computers and the Humanities*, 35(3): 315–31.
- Hope, J. and Witmore, M. (2010). The hundredth psalm to the tune of ‘Green Sleeves’: digital approaches to the language of genre. *Shakespeare Quarterly*, 61(3): 357–90.
- Jamal, N., Mohd, M., and Noah, S. A. (2012). Poetry classification using support vector machines. *Journal of Computer Science*, 8(9): 1441–6.
- Jockers, M. and Witten, D. M. (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, 25(2): 215–23.
- Jockers, M. (2013). *Macroanalysis: Digital Methods and Literary History*. Champaign, IL: University of Illinois Press.
- Kennedy, G. A. (1994). *A New History of Classical Rhetoric*. Princeton, NJ: Princeton University Press.
- Kumar, V. and Minz, S. (2014). Poem classification using machine learning approach. In *Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012)*, December 28–30, 2012, Jaipur, India, pp. 675–82. https://link.springer.com/chapter/10.1007/978-81-322-1602-5_72 (accessed 30 August 2017).
- Long, H. and So, R. J. (2016). Literary pattern recognition: modernism between close reading and machine learning. *Critical Inquiry*, 42(2): 235–67.
- Lorang, E., Soh, L. K., Datla, M. V., and Kulwicki, S. (2015). Developing an image-based classifier for detecting poetic content in historic newspaper collections. *D-Lib Magazine*, 21: 7–8. <http://www.dlib.org/dlib/july15/lorang/07lorang.html> (accessed 30 August 2017).
- Malmi, E., Takala, P., Toivonen, H., Raiko, T., and Gionis, A. (2016). DopeLearning: a computational approach to rap lyrics generation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco*, pp. 195–203. <http://dl.acm.org/citation.cfm?doid=2939672.2939679> (accessed 30 August 2017).
- Mayavavages, E., De Gussem, J., Saelemans, W., and Kestemont, M. (2017). Assessing the stylistic properties of neurally generated text in authorship attribution. In *Proceedings of the Workshop on Stylistic Variation, Copenhagen*, pp. 116–125. <http://aclweb.org/anthology/W17-4914> (accessed 4 June 2018).
- Marriott, I. (1979). The authorship of the *Historia Augusta*: two computer studies. *Journal of Roman Studies*, 69: 65–77.
- Matias, J. N. (2010). *Swift-Speare: Statistical Poetry*. <http://natematias.com/portfolio/DesignArt/Swift-SpeareStatisticalP.html> (accessed 30 August 2017).
- Mayer, R. (2005). The impracticability of Latin Kunstprosa. In Reinhardt, T., Lapidge, M., and Adams, J. N. (eds), *Aspects of the Language of Latin Prose. Proceedings of the British Academy*, vol. 129. Oxford: Oxford University Press, pp. 195–210.
- Moretti, F. (2013). *Distant Reading*. London: Verso.
- Morton, A. Q. and Winspear, A. D. (1971). *It’s Greek to the Computer*. Montreal: Harvest House.
- Mosteller, F. and Wallace, D. L. (1964). *Inference and Disputed Authorship: The Federalist*. Reading, MA: Addison-Wesley.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–30.
- Spevak, O. (2010). *Constituent Order in Classical Latin Prose*. Amsterdam: John Benjamins Publishing Company.
- Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3): 538–56.
- Stover, J., Winter, Y., Koppel, M., and Kestemont, M. (2016). Computational authorship verification method attributes a new work to a major 2nd century African author. *Journal of the Association for Information Science and Technology*, 67(1): 239–42.
- Tizhoosh, H. R. and Dara, R. A. (2006). On poem recognition. *Pattern Analysis and Applications*, 9(4): 325–38.
- Tizhoosh, H. R., Sahba, F., and Dara, R. (2008). Poetic features for poem recognition: a comparative

study. *Journal of Pattern Recognition Research*, 3(1): 24–39.

Vickers, B. (2004). *Shakespeare, Co-author: A Historical Study of Five Collaborative Plays*. Oxford: Oxford University Press.

Notes

- 1 Present address: 4D Path, Inc., Newton, MA, USA.
- 2 The authors are listed alphabetically, and the order of the author list does not reflect relative contributions to the work reported.
- 3 For instance, *Pede certo* is a tool for computational scansion of Latin dactylic hexameter (the meter of epic poetry) and elegiacs developed by the Università di Udine and the *Musisque Deoque* digital archive (<http://www.pedecerto.eu/>).
- 4 See <https://www.wolfram.com/mathematica/new-in-10/highly-automated-machine-learning/determine-if-a-text-is-prose-or-poetry.html>.

- 5 See, for instance, the ongoing development of the Classical Language Toolkit (CLTK), an extension of the Python NLTK library to Latin and Greek and, in due course, other ancient languages (www.cltk.org).
- 6 For example, *Allen and Greenough's New Latin Grammar*, which is available electronically through the Perseus Project (<http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.04.0001&redirect=true>). For the sake of completeness, we also included several archaic forms and alternative spellings of inflected forms.
- 7 We selected a diverse subset of prose and verse texts from the full corpus (Horace's *Odes* 1, Ovid's *Metamorphoses* 1, Plautus' *Amphitruo*, Vergil's *Aeneid* 1, Caesar's *De Bello Gallico* 1, Cicero's *Pro Archia*, and Vitruvius' *De Architectura*) and partitioned them into chunks of 500 words each, with any remaining material set aside. We then classified the chunks by genre using the full feature set and the same workflow as for the main experiments. The mean cross-fold accuracy was >90% as were the overall accuracy and *F1* score.

Large-scale quantitative profiling of the Old English verse tradition

Leonard Neidorf^{1,7}, Madison S. Krieger^{2,7*}, Michelle Yakubek^{3,4}, Pramit Chaudhuri⁵ and Joseph P. Dexter^{6*}

The corpus of Old English verse is an indispensable source for scholars of the Indo-European tradition, early Germanic culture and English literary history. Although it has been the focus of sustained literary scholarship for over two centuries, Old English poetry has not been subjected to corpus-wide computational profiling, in part because of the sparseness and extreme fragmentation of the surviving material. Here we report a detailed quantitative analysis of the whole corpus that considers a broad range of features reflective of sound, metre and diction. This integrated examination of fine-grained features enabled us to identify salient stylistic patterns, despite the inherent limitations of the corpus. In particular, we provide quantitative evidence consistent with the unitary authorship of *Beowulf* and the Cynewulfian authorship of *Andreas*, shedding light on two longstanding questions in Old English philology. Our results demonstrate the usefulness of high-dimensional stylometric profiling for fragmentary literary traditions and lay the foundation for future studies of the cultural evolution of English literature.

Composed between roughly 600 and 1100, Old English literature represents the earliest phase of literary production in English. Although it is assumed that most works of Old English literature have not survived, the remainder nevertheless encompass not only a broad time period but also multiple streams of influence—Germanic, Christian, and classical Greek and Roman—as well as diverse genres such as heroic poetry, riddles and biblical works¹. This rich corpus also contains one of the masterpieces of English literature—the epic poem *Beowulf*. Both for its historical importance and its aesthetic merit, Old English literature has attracted the attention of generations of researchers and creative writers, including W. H. Auden, Ezra Pound, Seamus Heaney and J. R. R. Tolkien^{2,3}.

Within Old English literature, the extant corpus of poetry is relatively small; it comprises around 350 texts, of which over 300 are shorter than 1,000 words in length (Supplementary Fig. 1). The poems are preserved in manuscript copies that provide no direct information about the context in which they originated. Moreover, damage to these manuscripts has frequently resulted in the loss of text, and rendered many poems more or less incomplete. The sparseness and fragmentation of the corpus poses a serious challenge for literary study of the tradition. For instance, it is almost impossible to know how representative, original or popular a particular literary feature might have been when we possess so few comparanda, the authorship or date of extant works is often unknown or uncertain, and the compositional technique of Old English poets is

similarly mysterious. Additional difficulties arise due to uncertainties of register, genre and dialect, which complicate efforts to relate literary works to particular chronological or geographical contexts⁴. Lacking the extensive corpora and contextual evidence that are taken for granted in the study of modern literatures, Old English scholars face considerable difficulties when dealing with questions of literary history. Some scholars have even suggested that the surviving materials are insufficient for meaningful conclusions to be drawn on the basis of linguistic analysis⁵.

One approach to these problems is to extract more information from the material we already have; rather than examining larger and therefore less frequent components of the literature, such as characters or scenes, we can focus on much smaller units, ranging from individual phrases to word segments and even pauses. A benefit of analysing smaller features is that they are necessarily numerous even within a sparse corpus. By combining attention to multiple features of this kind, it is possible to create a high-dimensional profile of a text, or part of a text, in relation to all others in the corpus. Although manual counting of individual small features may be feasible, the generation of high-dimensional profiles generally requires the application of computational techniques. Such techniques have not been employed extensively in the study of Old English literature compared with modern English or even other pre-modern traditions such as Latin^{6–9}. Where computation has been brought to bear on Old English texts, the research has generally been limited to a small set of literary features or a handful of specific works^{10–15}. The most significant application of modern stylometric techniques to Old English verse has been the development of ‘lexomics’ by Drout et al., which involves the use of vocabulary frequency data and hierarchical clustering to discern literary similarities^{12,15}. Lexomic methods have been applied to several important problems, including profiling stylistic differences across *Beowulf* and works associated with Cynewulf (the first author to whom multiple English poems can be attributed)^{12,15}. Our methodology complements and extends this earlier work in three principal ways: (1) the use of non-lexical features, especially sense-pauses and metre; (2) attention to specialized word usage, in particular rare nominal compounds; and (3) adaptation of clustering techniques to focus on sequences of characters rather than whole words.

Here, we report a large-scale computational analysis of the entire Old English verse corpus. Our central innovation is to extract information on a wide range of fine-grained features—covering aspects of sound, metre and diction—to discern meaningful stylometric patterns within the corpus as they relate to questions of authorship

¹Department of English, Nanjing University, Nanjing, China. ²Program for Evolutionary Dynamics, Harvard University, Cambridge, MA, USA.

³Research Science Institute, Center for Excellence in Education, McLean, VA, USA. ⁴Texas Academy of Mathematics and Science, Denton, TX, USA.

⁵Department of Classics, University of Texas at Austin, Austin, TX, USA. ⁶Department of Systems Biology, Harvard Medical School, Boston, MA, USA.

⁷These authors contributed equally: Leonard Neidorf, Madison S. Krieger. *e-mail: mkrieger@fas.harvard.edu; jdexter@fas.harvard.edu

and literary resemblance. Additionally, we introduce a variety of computational tools tailored to the specifics of Old English, which in turn may be valuable in the analysis of other literary and linguistic traditions. We use our corpus-wide profiling to address two longstanding questions in the study of Old English literature: whether *Beowulf* is a unified work of a single author or a combination of multiple texts^{16,17}, and whether the anonymous work *Andreas* was written by the poet Cynewulf^{18–20}. We show that several orthogonal stylistic metrics do not differ between possible partitions of *Beowulf*, which is consistent with the hypothesis that portions of the poem were not produced separately, or, if they were, that the styles are remarkably uniform. Although this uniformity cannot adjudicate definitively between single or multiple authorship, it militates against a view of the work either as constructed from chronologically disparate poems or as markedly shaped by scribal intervention. Our results also show strong similarity between *Andreas* and other works signed by Cynewulf. Our approach has implications not only for the practice of literary criticism but also for the study of cultural evolution, by generating data on properties of language that probably evolved from this early tradition through Middle and Modern English²¹. While computational analyses cannot definitively resolve longstanding problems arising from a dearth of empirical evidence and theoretical disagreements about cultural production, nevertheless, they do offer additional, quantifiable data that affect the plausibility of various critical hypotheses.

Functional *n*-grams are short (typically syllable-length) substrings of natural language text (for example, the substring ‘ab’ in the sentence ‘Abel elaborated about his intentions.’), which have proven useful in previous analyses of both English and Latin literary style, and for authorship attribution, as works by the same author tend to have similar phonetic profiles^{9,22–25}. In verse corpora, patterns of functional *n*-gram usage can reflect poetic sound play and aural effects. To identify phonetically distinctive poems within the Old English corpus, we computed for each text:

$$\sum_{i=1}^5 |f_{i,t} - f_{i,c}|$$

where $f_{i,t}$ denotes the frequency of the *i*th most common *n*-gram in the text, and $f_{i,c}$ denotes the corpus-wide frequency of that *n*-gram. Figure 1 shows a plot of this metric against text length for functional trigrams. Unsurprisingly, numerous short poems appear to have patterns of functional *n*-gram usage that differ from the bulk corpus. However, of greater interest is that 3 longer texts (each longer than 125 verses) exhibit unusual patterns of functional trigram usage relative to other texts of comparable length. Profiling of functional bigrams and four-grams similarly identified these same three texts—*Widsith*, *Psalm 118* and *Maxims II*—as anomalous (Supplementary Fig. 2). In other words, these three texts exhibit pronounced deviations from phonetic norms that are otherwise relatively homogeneous throughout the corpus of Old English poetry.

These results prompted us to consider why the phonetics of those three texts should appear distinct from the rest of the corpus. In the case of *Widsith*, the anomaly is likely to be attributed to its preponderance of proper names, which are clustered in 3 lengthy catalogues (lines 18–35, 57–87 and 112–124) that might well have circulated orally before the poem’s composition²⁶. If proper names are the cause, the anomalous phonetics of *Widsith* might be an epiphenomenal reflection of a broader phonetic division between the lexicon and the onomasticon of Old English. *Maxims II* shares with *Widsith* the strong possibility that its author drew on pre-existing material, consisting as it does of gnomic statements that could have circulated in smaller or larger catalogues outside the poem. As such, it is plausible that its anomalous *n*-gram profile reflects phonetic differences between the archaic constituent material of *Maxims II*

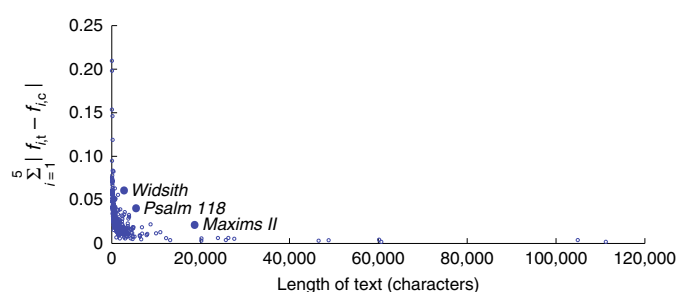


Fig. 1 | Corpus-wide phonetic profiling of literature. Plot of the cumulative difference in functional trigram frequency (for the five most common trigrams) against text length. Each dot denotes one text. Three anomalous texts are highlighted and labelled.

and the later linguistic material of which the bulk of the Old English corpus is comprised. *Psalm 118* is peculiar less for its content than for its aberrant metrics and late prosaic vocabulary. There is frequent lexical and syntactic repetition in *Psalm 118*, whereas *Widsith* and *Maxims II* exhibit a far greater degree of structural repetition. As a close translation of a Latin source that might have originated as an interlinear gloss, *Psalm 118* shares with *Widsith* and *Maxims II* the more essential characteristic that its poet’s linguistic freedom was exceptionally constrained by his literary project. Our tests suggest that, under normal conditions, Old English poets generated works that were homogeneous in terms of their phonetic profile. However, this homogeneity was disturbed when a particular literary agenda strongly influenced a poet’s diction or source use.

In addition to analysing the phonetics of the Old English corpus in its entirety, we also sought to address longstanding questions regarding particular texts, beginning with *Beowulf*. Scholarship on *Beowulf* has long entertained debate as to whether the poem is a product of unitary or composite authorship. During the nineteenth century, many prominent scholars subscribed to a theory of composite authorship, which held that *Beowulf* consisted of various pagan lays joined together by Christian editors and interpolators¹⁶. By the middle of the twentieth century, this view possessed few adherents on account of demonstrations by Klaeber²⁷ and Tolkien² that a coherent Christian perspective pervades the entire poem. Literary critics working in the immediate aftermath of these studies thus tended to premise their work on the assumption that *Beowulf* is the masterwork of a single poet^{28,29}. However, theories of composite authorship continued to be propounded throughout the twentieth century, with several scholars arguing that *Beowulf* was put together by a scribal editor who combined two distinct texts: one containing the hero’s fights with Grendel and his mother, and the other containing the hero’s fight with the dragon^{30–32}. It has also been argued that scribal interference in the textual transmission of *Beowulf* might have been sufficiently pervasive to render it an essentially composite work³³. Yet, in the most recent and comprehensive study on the dating and authorship of *Beowulf*, Neidorf¹⁷ adduced a wide range of lexical, metrical, stylistic and palaeographical evidence in support of the contention that the extant manuscript of *Beowulf* faithfully preserves the unitary creation of one poet who composed around the year 700. Here, we offer multiple, orthogonal pieces of quantitative evidence consistent with Neidorf’s view.

As noted above, quantitative analysis of stylistic homogeneity in *Beowulf* has tended to focus on word-based features. To investigate the question further, we devised a broad-spectrum feature set that reflects versification, metre and an aspect of diction (nominal compounds) of particular importance in Old English verse. We first considered sense-pauses, which are breaks in speech typically denoted by any punctuation mark other than a comma. Although sense-pause analysis has not been undertaken previously for Old

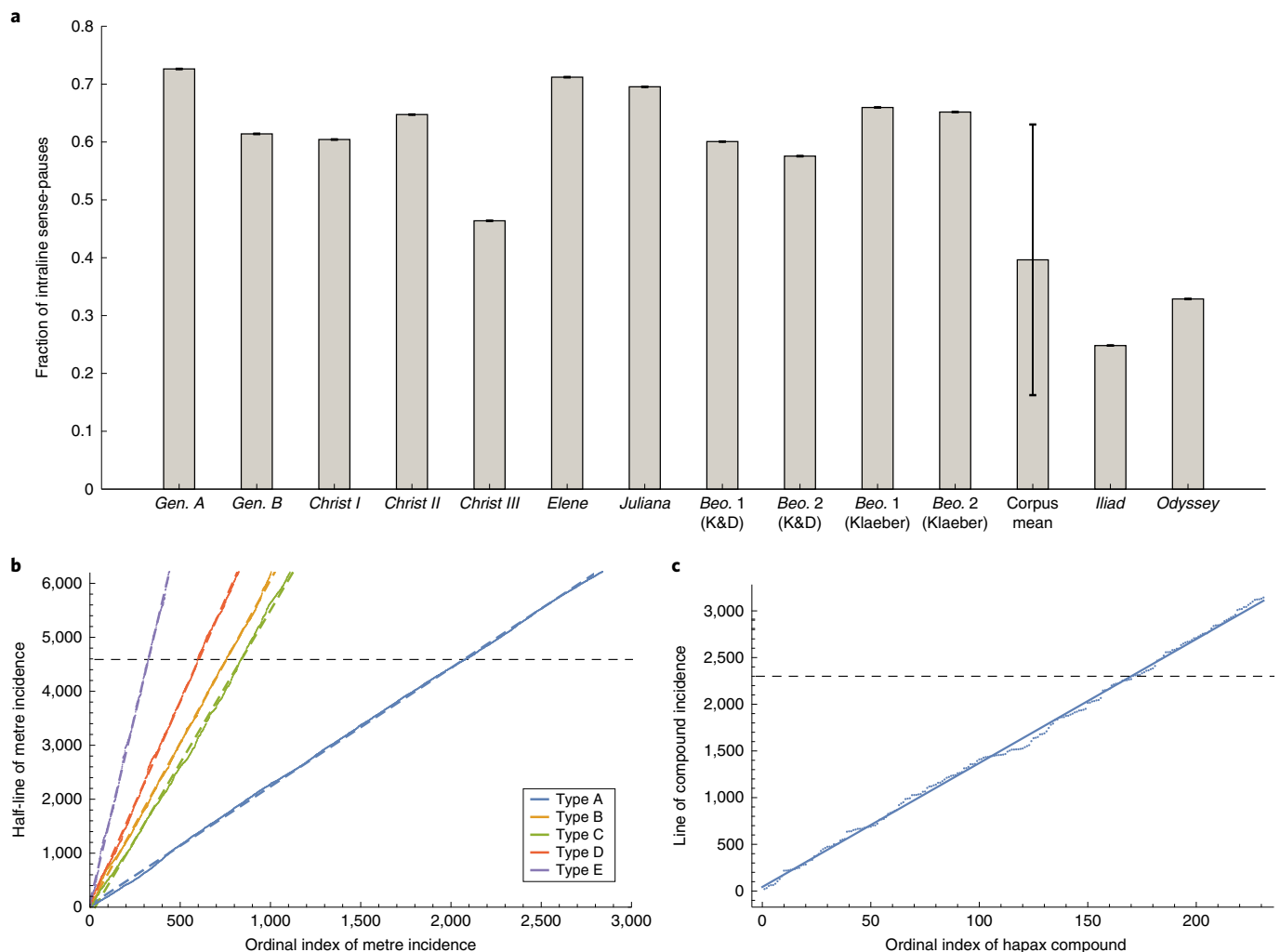


Fig. 2 | Stylistic homogeneity of *Beowulf*. **a**, Ratio of intraline to total sense-pauses for the partition of *Beowulf*, the mean of all texts in the Old English verse corpus, some salient individual texts, and Homer's *Iliad* and *Odyssey*. The error bar for the corpus mean denotes one s.d. of the ratios for all of the texts. *Beo.*, *Beowulf*; *Gen.*, *Genesis*; K&D, Krapp and Dobbie. **b**, Rate of use of different metres (types A, B, C, D and E) in *Beowulf*. The coloured diagonal lines are linear fits. The dotted horizontal line indicates line 2,300 of the text. **c**, Rate of use of hapax compounds in *Beowulf*. The blue line is a linear fit. The dotted horizontal line indicates line 2,300 of the text.

English literature, it has been applied to questions of stylistic evolution in other traditions. For instance, Fitch³⁴ demonstrated that the ratio of intraline to total sense-pauses is a reliable marker of relative chronology for the tragedies of Sophocles, Seneca and Shakespeare, perhaps because frequent inclusion of sense-pauses not coincident with line breaks reflects a more confident and mature poetic style.

Theories of composite authorship differ as to the exact division between the component poems, but most suggestions (for example, refs. ^{16,17}) cluster around line 2,300, which is not long after the scribal hand changes in the manuscript (in the middle of line 1,939)³². As such, we computed the ratio of intraline to total sense-pauses for lines 1–2,300 and 2,301–end of Krapp and Dobbie's edition of *Beowulf*, along with the corpus-wide average. We calculated that the ratios for lines 1–2,300 and 2,301–end are within 4% of each other (Fig. 2a). As is typical for pre-modern texts, there is no punctuation in the extant Old English manuscripts. However, the editorial judgements about where punctuation is required are guided by various metrical and syntactic regularities, such as those codified in Kuhn's laws, which reliably indicate where clauses begin and end in Old English poetry^{35,36}. To account for the remaining freedom in editorial practice regarding punctuation, we analysed another

text of *Beowulf* (edited by Klaeber and revised by Bjork, Fulk and Niles). We found that although the absolute value of the ratios varies between the two editions, the relative difference between lines 1–2,300 and 2,301–end is small in both cases. This comparison suggests that while editorial policies may differ, their consistent application ultimately does not obscure the stylistic regularities in a given poem. The consistency in the handling of intraline sense-pauses across both sections of *Beowulf*, in both editions, therefore provides support for the stylistic unity of the poem.

We sought to corroborate the results of the sense-pause analysis of *Beowulf* through comparison with other Old English poems and with ancient Greek epic. *Genesis*—one of the longest extant Old English poems—is known to be the work of multiple authors; it consists of a later poem, called *Genesis B*, which is approximately 600 lines long and is embedded within the remaining 2,300 or so lines of the older main poem, *Genesis A*. Differences between the two poems have previously been identified using other conventional and quantitative techniques¹². In our research, we found a marked difference in the intraline-to-total sense-pause ratio between *Genesis A* and *B* (Fig. 2a), suggesting that sense-pause analysis can distinguish between passages of Old English verse about

similar subject matter but composed by different poets. Likewise, the ratio differs between all three *Christ* poems (*Christ I–III*), which are widely held to have been composed by multiple authors. In contrast, we find that the ratio is consistent between *Elene* and *Juliana*, both of which are signed Cynewulfian poems. In aggregate, sense-pause differences are significantly higher in the *Genesis* and *Christ* group than in the *Cynewulf* and *Beowulf* group (two-tailed *t*-test, $t(5)=2.94$; $P=0.0322$; Cohen's $d=2.25$; 95% confidence interval (CI) = 0.013 to 0.194).

Like *Beowulf*, the Greek epics *Iliad* and *Odyssey* have also generated much debate about their authorship and composition. Conventionally attributed to a single author—Homer—both works nevertheless clearly originate in a long oral tradition and show signs of considerable evolution in the course of their transmission history, including the possible influence of written versions^{37,38}. Since the two Homeric epics have numerous features in common, we hypothesized that they might also have a similar pattern of sense-pauses. However, as shown in Fig. 2a, the *Odyssey* has a higher proportion of intraline sense-pauses relative to the *Iliad*. This difference suggests a slight change of compositional practice between the two Greek poems, whether due to a single poet's stylistic evolution or natural variation across the oral tradition. Had the two parts of *Beowulf* shown a similar or greater disparity in the sense-pause data when compared with the *Iliad* and the *Odyssey*, this might have supported the view that two different poems had been conjoined. However, as it stands, the comparative uniformity of the data suggests that the compositional practice of both parts was the same, at least with respect to sense-pauses.

We then examined the metre of *Beowulf*. We used a scansion devised by Sievers³⁹, which categorizes half-lines into five major sound patterns denoted as types A, B, C, D and E. We investigated both the total frequency of the five verse-types and their sequence within *Beowulf*. Strikingly, we found that the usage rate of each type of metre remains linear across the entirety of *Beowulf* (Pearson's $r(2,860)=0.998$; $P<0.001$ for type A; $r(1,008)=0.997$; $P<0.001$ for type B; $r(1,241)=0.998$; $P<0.001$ for type C; $r(826)=0.997$; $P<0.001$ for type D; $r(445)=0.995$; $P<0.001$ for type E), with no discernible shift near line 2,300 and no differences in the frequencies of any particular metrical type (Fig. 2b). To quantify this effect, we computed the difference in slope between the two sections of 1,000 randomly chosen partitions of *Beowulf*; for 4 of the 5 metre types, the mean difference across the partitions is greater than the difference between lines 1–2,300 and 2,301–end (mean = 0.148; s.d. = 0.063 versus 0.0498 for type A; mean = 0.715; s.d. = 0.884 versus 0.331 for type B; mean = 0.717; s.d. = 0.389 versus 0.492 for type C; mean = 0.524; s.d. = 0.697 versus 0.750 for type D; mean = 1.56; s.d. = 1.75 versus 0.575 for type E).

Finally, we considered the distribution of nominal compounds in *Beowulf* and across the Old English verse corpus. Nominal compounds, which are words formed by combining two nouns, are a particularly important aspect of Old English poetry⁴⁰. Examples in Old English include *hron-rad* (whale-road), referring to the sea, and *ban-hus* (bone-house) for the human body. It is generally believed that the number and inventiveness of compounds in a poem is an important marker of literary creativity in the Old English tradition^{1,40}. To generate a list of compound words, we identified all entries in the Bosworth–Toller dictionary that are connected with a hyphen and that consist of two separate headwords (hyphens do not appear in native Old English texts)⁴¹. Supplementary Fig. 3 shows the distribution of compound words by frequency of occurrence. For our initial analysis, we considered inter-authorial differences in the usage of hapax legomena compound words (that is, compounds that appear only once in the entire poetic corpus). The rate of usage of hapax compounds can be very different between authors, as illustrated in Supplementary Fig. 4 for two of the longest extant poems (*Genesis* and *Exodus*). As discussed above, *Genesis* is known to be a

composite work. We partitioned *Genesis A* into two random sections that are of comparable length to *Genesis B* and analysed the rate of hapax usage. Linear fits to the data for both sections of *Genesis A* have very similar slopes and differ from the fit to the *Genesis B* data (Supplementary Fig. 4a). In contrast, *Exodus*—the unitary authorship of which has never been in dispute—shows clear homogeneity when analysed in the same way (Supplementary Fig. 4b).

Taken together, these results demonstrate that nominal compounds are an effective metric for Old English stylistic and attribution studies. We therefore constructed a profile of hapax compounds across the whole of *Beowulf* (Fig. 2c). This profile revealed that the rate of compound usage is linear throughout the poem (Pearson's $r(229)=0.992$; $P<0.001$), with no change in slope observed around line 2,300. The difference in slope between lines 1–2,300 and 2,301–end is 1.50 (mean = 1.42; s.d. = 1.02 for 1,000 random partitions). The small nonlinearity evident around line 1,500 corresponds to *Beowulf*'s fight with Grendel's mother, which is known to be particularly rich in compound words and other distinctive linguistic features. Accordingly, our analysis of nominal compounds provides further evidence (orthogonal to the sense-pause and metrical data) for the stylistic homogeneity of *Beowulf*.

Our other major results concern a collection of poems written by an author called Cynewulf, or by a broader 'Cynewulfian school'. Four Old English poems—*Elene*, *Juliana*, *Christ II* and *Fates of the Apostles*—conclude with epilogues that ascribe their composition to an otherwise unknown individual named Cynewulf. Many scholars, perceiving stylistic or thematic affinities between these signed works and other anonymous poems, have sought to expand Cynewulf's corpus to include such works as *Andreas*, *Guthlac A/B*, *Christ I/III*, *Judith*, *The Phoenix* and *The Dream of the Rood*, among other poems^{42–44}. While once considered products of Cynewulf's own hand, these poems are now more commonly regarded as products of a Cynewulfian school of poetry, if they are believed to possess any meaningful connections to his work at all⁴⁵. The majority of scholars in the past half-century consider Cynewulf to be the author of only the four signed poems, although some have maintained that either *Guthlac B*, *Andreas* or both should be included in his corpus as well, in part based on computational stylometric analysis^{12,46}. In addition, whether Cynewulf should even be regarded as the author of the four poems bearing his signature has been questioned, since it is theoretically possible that Cynewulf added his epilogues to poems that other authors originally composed^{20,47}. Our tests assuage such doubts by identifying a strong degree of stylistic homogeneity among three of the four signed works of Cynewulf. This homogeneity supports the longstanding assumption that one author composed at least three, and possibly all four, of the poems in question. We also find compelling evidence for an association between *Andreas* and Cynewulf's poetry, which might indicate—in contrast with current opinion—that Cynewulf composed this poem as well.

We first compared the usage of hapax compounds across ten Old English poems, including three control texts not by Cynewulf (*Beowulf*, *Exodus* and *Christ and Satan*), the four signed Cynewulf poems and three poems often associated with Cynewulf (*Andreas*, *Guthlac B* and *The Phoenix*) (Supplementary Fig. 4c). The three control poems, which are thought to be by different authors writing during different periods in Anglo-Saxon history, unsurprisingly show distinct patterns of compound usage. However, the signed Cynewulf poems appear similar both to each other (although *Christ II* shows less affiliation with the other works) and to *Andreas*.

This result prompted us to examine the similarity of *Andreas* to the signed poems of Cynewulf on the basis of a broader range of nominal compounds beyond hapax legomena. In Old English, multiple compounds could denote a single object or concept. There are at least 17 completely distinct compound words in the poetic corpus that denote 'the sea', for example, and 11 compounds meaning 'warrior'^{48,49}. Therefore, an author's particular choice of compound



Fig. 3 | Use of nominal compounds is similar between Cynewulf and Andreas. Distribution of non-unique compounds in six poems either signed by Cynewulf (blue) or of possible Cynewulfian authorship (red) and in *Beowulf* (grey). The size of each filled circle indicates the number of compounds shared by the corresponding pair of texts. The dotted open circles indicate the expected size if all compounds were distributed at random. The circle in the bottom right is for comparison of *Beowulf* 1–2,300 and *Beowulf* 2,301–end.

might reflect a variety of factors: nuance in meaning, literary influence or other linguistic considerations. Accordingly, the usage of compounds forms an important part of the stylistic profile of an Old English author. We performed a large-scale analysis of non-hapax compounds (excluding only *wuldorcyning*, *heofoncyning* and *heofonrice* ('wonder-king', 'heaven-king' and 'heaven-kingdom', respectively), which occur with extremely high frequency throughout much of the Old English religious poetry) in the verse corpus (Fig. 3). Each solid circle in Fig. 3 denotes the degree of correlation between the two indicated texts, compared with a random distribution of compound words based on their overall frequency and the lengths of the individual poems (dotted circles). By this measure, most of the signed works of Cynewulf are strongly correlated. However, *Christ II* is close to naively correlated, perhaps due to an absence of the compounds that are typically used in the hagiographical poems (*Elene* and *Juliana*) to express the divine relationship between the saint and God (for example, *mundbyrd* ('suffrage/aid')). By the same measure, *Andreas* is strongly correlated with the signed poems of Cynewulf, in agreement with our analysis of hapax compounds (Supplementary Fig. 4). Supporting these observations, at least one of the Cynewulf/Cynewulf (blue circles), Cynewulf/*Andreas* (top line of red circles) and Cynewulf/other (remaining red circles) comparison groups is significantly different from the others (one-way analysis of variance, $F(2,18)=5.73$; $P=0.0119$). There is no significant pairwise difference between Cynewulf/Cynewulf and Cynewulf/*Andreas* ($Q(18)=0.653$; $P=0.890$; Cohen's $d=-0.244$; 95% CI= -0.952 to 1.37), but there is a significant difference between Cynewulf/Cynewulf

and Cynewulf/other ($Q(18)=3.75$; $P=0.0410$; Cohen's $d=1.32$; 95% CI= -1.86 to -0.0354) and between Cynewulf/*Andreas* and Cynewulf/other ($Q(18)=3.98$; $P=0.0294$; Cohen's $d=2.04$; 95% CI= -2.21 to -0.108 ; all values are from post-hoc Tukey–Kramer tests). Additionally, we observe that *Beowulf* is self-correlated when partitioned into lines 1–2,300 and 2,301–end, which provides further support for unitary composition.

Finally, we used hierarchical agglomerative clustering to investigate the possible association of *Andreas* with the signed poems of Cynewulf on the basis of functional n -gram frequencies, which are often used for authorship attribution studies involving literary texts written in Modern English^{22,24}. As described in detail in the Methods, we computed the frequencies of the 25 most common trigrams (based on the corpus-wide frequency) in the 50 longest poems, with *Beowulf* partitioned into 2 parts as usual. We used hierarchical agglomerative clustering with this feature set to construct the dendrogram shown in Fig. 4. In line with our other studies, *Beowulf* lines 1–2,300 and 2,301–end cluster together. Furthermore, we find that *Andreas* clusters next to *Elene* and in close proximity to *Juliana*, *Fates of the Apostles* and *Christ I/II/III*. Also in this cluster is *Guthlac A* and *B*, the latter of which Drout et al. associated with the works of Cynewulf based on a clustering analysis with word-level features and a small subset of the Old English verse corpus¹². To investigate the robustness of these observations, we repeated the clustering with bigrams and four-grams and found that key aspects of the dendrogram structure, including the side-by-side positioning of the two parts of *Beowulf*, and the positioning of *Andreas* next to *Elene* and in close proximity to at least two other signed

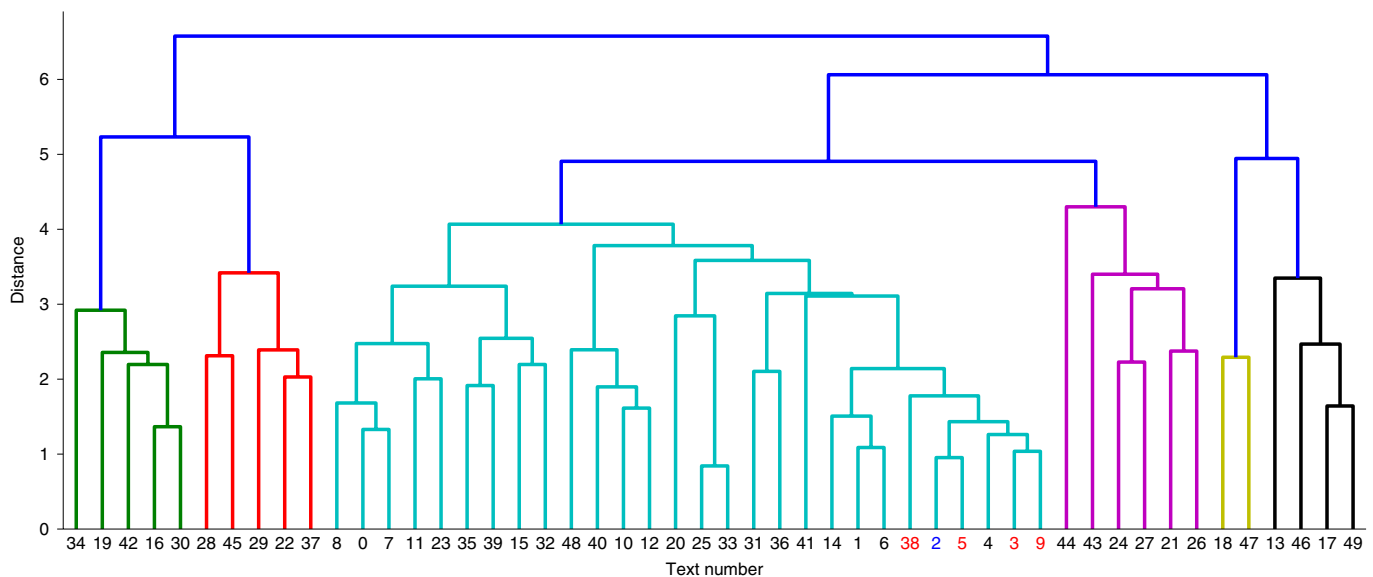


Fig. 4 | Andreas clusters with the signed Cynewulfian poems. Dendrogram generated from hierarchical agglomerative clustering of the 50 longest poems, with high-frequency functional trigrams as features. The text numbers are given in Supplementary Table 1; for reference, *Andreas* is text 2 (blue) and the Cynewulfian poems are all red (*Christ*, 3; *Elene*, 5; *Juliana*, 9; and *Fates of the Apostles*, 38). *Beowulf* 1–2,300 is 1, and *Beowulf* 2,301–end is 6.

Cynewulfian poems, were preserved in both cases (Supplementary Fig. 5). Our results, obtained using unsupervised learning and a type of feature (high-frequency functional n -grams) well-established in the attribution literature, thus corroborate the stylistic association between Cynewulf and *Andreas* that we identified through analysis of nominal compounds.

The tests we have conducted indicate some ways in which quantitative profiling of the Old English verse tradition can help to answer or raise questions of considerable interest to researchers. With regard to *Beowulf*, our tests tilt the scales of probability between hypotheses that are currently in competition. In contrast, with Cynewulf, our tests encourage scholars to reconsider a possibility that has not been seriously entertained in the past half-century. Our evidence for the stylistic homogeneity of *Beowulf* does not prove that the poem is the work of one individual, but it substantially enhances the probability of unitary authorship, while presenting serious obstacles to those who would advocate for composite authorship or scribal recomposition. Our evidence for the extraordinary affinities between the language of *Andreas* and the language of the signed works of Cynewulf similarly does not prove that *Andreas* was composed by Cynewulf, but it demands that this possibility be explored further in future studies. Orchard, noticing many formulaic expressions shared between *Andreas* and the works of Cynewulf, interpreted the overlap as an indication that the *Andreas* poet read the works of Cynewulf and borrowed extensively from them¹⁹. Given the lack of decisive evidence, we must acknowledge the possibility, both for *Beowulf* and *Andreas*, that some combination of generic constraints and highly skilled imitation might account for the patterns observed. In each case, however, the most economical explanation of the data is similar: unitary composition of *Beowulf* and Cynewulfian authorship of *Andreas*. Furthermore, in view of the fact that *Andreas* immediately precedes *Fates of the Apostles* in the Vercelli Book (the manuscript in which these two poems are preserved), we might tentatively regard Cynewulf's signature at the end of the latter as a claim to authorship of the former as well.

Our results show the utility of taking a wide range of quantitative approaches to the study of a literary corpus, from simple frequency counts to machine learning. However, crucial to the success of any large-scale profiling is the selection of features used to characterize

the corpus⁵⁰. In this case, the variety of features complements and enhances the more established focus on word usage and distribution, incorporating in addition phonetic, formulaic, rhythmic and metrical elements. In doing so, we exploit features that are known to play an important role in the specific tradition (for example, nominal compounds), as well as validate the extension of features that have proven useful for studying traditions in other languages (for example, functional n -grams and sense-pauses) to Old English^{9,23,34}. In our analysis of Cynewulf, we show that a corpus-specific feature (nominal compounds) can be combined with a general-purpose stylistic technique (unsupervised learning with character n -grams) to provide broad-based support for the Cynewulfian authorship of *Andreas*. Moreover, the quantitative tests designed to analyse nominal compounds might be profitably applied in the future to other languages and traditions where aspects of word formation allow for the free combination of simpler lexical items into larger, often unique units, such as agglutinative and polysynthetic languages⁵¹. In summary, our diverse combination of methods and features constitutes an effective response to the challenges posed by sparse corpora. In particular, the computational analysis of many microscopic features yields results that either cannot be obtained using conventional critical methods or can only be obtained with great difficulty.

Our approach provides a model applicable to other literary traditions. Although potentially useful for the analysis of any corpus of literature, the techniques described here offer a particular advantage for the study of corpora posing similar challenges as Old English poetry, such as other medieval traditions including Old Norse, Old Irish and Old French^{52–54}. These languages exhibit many characteristics shared with Old English and are hence especially amenable to the same methods. However, all pre-modern literary traditions suffer to a greater or lesser extent from the problem of text loss, and hence sparse corpora—a situation compounded by the frequent lack of contextual information about the date or authorship of works. Our study suggests some general ways of overcoming or circumventing these challenges, and of finding data that can shed light on both work-specific and corpus-wide questions. Generating quantitative profiles for multiple literary traditions would also represent an initial step towards a quantitative analysis of literature

across cultures. Furthermore, in focusing on a pre-modern tradition—especially one that has seen relatively little computational research—our work broadens the digital humanities' predominant concern with modern literature, and lays the foundation for future diachronic profiling of the English literary tradition with substantial time depth⁵⁵.

Methods

Corpora and text processing. The texts of the Old English verse corpus were obtained from the University of Calgary's Online Corpus of Old English Poetry (OCOEP) in UTF-8 encoding (<http://www.oepoetry.ca/>), which preserves native Old English characters and contains character markings separating half-lines and full-lines, as well as different poems. Except in the following two cases, we used the complete, unaltered OCOEP for corpus-wide analyses: (1) before computing the corpus mean for sense-pauses (Fig. 2a), we aggregated related short texts (for example, the poems of the *Paris Psalter* and the *Meters of Boethius* into single files; and (2) we restricted the hierarchical clustering analysis to the 50 longest texts in the unaltered corpus, with the two partitions of *Beowulf* counted separately (Fig. 4 and Supplementary Fig. 5). Greek texts of the *Iliad* and *Odyssey* were obtained from the Tesseract Project, whose corpus is derived from the Perseus Digital Library (<http://www.perseus.tufts.edu/hopper/>).

Natural language processing. All natural language processing tasks were performed using Python 3.6.4.

Calculation of sense-pause frequency. Following the definition of Fitch for Greek, Latin and modern English poetry³⁴, we determined sense-pause frequencies by tabulation of punctuation marks other than commas (., ?, !, ;, :, (,), -, ', " and "). Any punctuation mark not coincident with a line break was considered to be an intraline sense-pause.

Metrical analysis. To supplement the OCOEP text file of the corpus for metrical analysis, we sought scansion of the longest poems, which were provided by G. Russom⁵⁶. We then identified for *Beowulf* the total frequency of each of the five verse-types, as well as the half-lines on which they occurred.

Identification of nominal compounds. We compiled a list of compound words from the set of hyphenated noun–noun headwords in the online Bosworth–Toller dictionary⁵⁷, excluding only one compound (*middangeard*, which means 'middle-land' or 'Earth' and inspired Tolkien's 'Middle Earth'). This compound is used with unusually high frequency (135 instances) and appears to have been used in a manner distinct from other poetic compounds throughout most of the Old English literary period. For each compound in the list, we identified the set of all poems in which that compound occurs, which was used to generate Fig. 2c and Supplementary Fig. 4. We also computed a measure of correlation between poems, defined relative to a random distribution of compound words according to a compound's frequency and the length of the poem. This random distribution of compound words was calculated 10,000 times per compound word to generate a well-defined distribution over the corpus for that word. For Fig. 3, we summed all of the compound words that appear in multiple poems, which quantifies the extent to which each pair of poems has shared compounds. The radius of each circle in Fig. 3 is the ratio of this number to the number predicted by the random distribution.

Hierarchical agglomerative clustering. To generate feature sets for clustering analysis, we determined the 25 most common functional bigrams, trigrams and four-grams in the Old English verse corpus and computed their frequency in the 50 longest poems (with *Beowulf* partitioned into lines 1–2,300 and 2,301–end). We used the scipy implementation of hierarchical agglomerative clustering with the Euclidean distance metric and Ward's linkage criterion to cluster those 50 texts. Dendrograms for the bigram and four-gram clustering are shown in Supplementary Fig. 5, and a dendrogram for the trigram clustering is shown in Fig. 4.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All datasets are freely and publicly available at <https://github.com/qcrist>.

Code availability

All custom code is freely and publicly available at <https://github.com/qcrist>.

Received: 23 November 2017; Accepted: 3 March 2019;

Published online: 08 April 2019

References

- Fulk, R. & Cain, C. A *History of Old English Literature* 2nd edn (Wiley-Blackwell, 2013).
- Tolkien, J. *Beowulf: the monsters and the critics*. *Proc. Br. Acad.* **22**, 245–295 (1936).
- Clark, D. & Perkins, N. *Anglo-Saxon Culture and the Modern Imagination* (D. S. Brewer, 2010).
- Biber, D. & Conrad, S. *Register, Genre, and Style* (Cambridge Univ. Press, 2009).
- Amos, A. C. *Linguistic Means of Determining the Dates of Old English Literary Texts* (Medieval Academy of America, 1980).
- Jockers, M. *Macroanalysis: Digital Methods and Literary History* (Univ. Illinois Press, 2013).
- Long, H. & So, R. Literary pattern recognition: modernism between close reading and machine learning. *Crit. Inq.* **42**, 235–267 (2016).
- Chaudhuri, P. & Dexter, J. P. Bioinformatics and classical literary study. *Journal of Data Mining & Digital Humanities Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages* <https://jdm.dh.episciences.org/paper/view?id=3807> (2017).
- Dexter, J. et al. Quantitative criticism of literary relationships. *Proc. Natl Acad. Sci. USA* **114**, E3195–E3204 (2017).
- Barquist, C. Phonological patterning in *Beowulf*. *Lit. Linguist. Comput.* **2**, 19–23 (1987).
- Barquist, C. & Shie, D. Computer analysis of alliteration in *Beowulf* using distinctive feature theory. *Lit. Linguist. Comput.* **6**, 274–280 (1991).
- Drout, M. D., Kahn, M. J., LeBlanc, M. D. & Nelson, C. Of dendrogrammar: lexicomic methods for analyzing relationships among Old English poems. *J. Eng. Ger. Philol.* **110**, 301–336 (2007).
- García, A. M. & Martín, J. C. Function words in authorship attribution studies. *Lit. Linguist. Comput.* **22**, 49–66 (2007).
- Gill, P., Swartz, T. & Treschow, M. A stylometric analysis of King Alfred's literary works. *J. Appl. Stat.* **34**, 1251–1258 (2007).
- Drout, M., Kisor, Y., Smith, L., Dennett, A. & Piirainen, N. *Beowulf Unlocked: New Evidence from Lexomic Analysis* (Palgrave Macmillan, 2016).
- Shippey, T. in *A Beowulf Handbook* (eds Bjork, R. & Niles, J.) 159–168 (Univ. Nebraska Press, 1998).
- Neidorf, L. *The Transmission of Beowulf* (Cornell Univ. Press, 2017).
- Bjork, R. *Cynewulf: Basic Readings* (Garland Publishing, 1996).
- Orchard, A. in *Anglo-Saxon Styles* (eds Karkov, C. & Brown, G.) 271–305 (State Univ. New York Press, 2003).
- Puskar, J. Questioning Cynewulf's claim of authorship. *Eng. Stud.* **92**, 1–19 (2011).
- Mesoudi, A. Pursuing Darwin's curious parallel: prospects for a science of cultural evolution. *Proc. Natl Acad. Sci. USA* **114**, 7853–7860 (2017).
- Grieve, J. Quantitative authorship attribution: an evaluation of techniques. *Lit. Linguist. Comput.* **22**, 251–269 (2007).
- Forstall, C., Jacobson, S. & Scheirer, W. Evidence of intertextuality: investigating Paul the Deacon's *Angustae Vitae*. *Lit. Linguist. Comput.* **26**, 285–296 (2011).
- Koppel, M., Schler, J. & Argamon, S. Computational methods in authorship attribution. *J. Assoc. Inf. Sci. Technol.* **60**, 9–26 (2009).
- Sapkota, U., Bethard, S., Montes, M. & Solorio, T. Not all character *n*-grams are created equal: a study in authorship attribution. In *Proc. 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 93–102 (Association for Computational Linguistics, 2015).
- Neidorf, L. The dating of *Widsith* and the study of Germanic antiquity. *Neophilologus* **97**, 165–183 (2013).
- Klaeber, F. *The Christian Elements in Beowulf* (Medieval Institute Publications, 1996).
- Brodeur, A. *The Art of Beowulf* (Univ. California Press, 1959).
- Irving, E. *A Reading of Beowulf* (Yale Univ. Press, 1968).
- Schücking, L. L. *Beowulfs Rückkehr: Eine Kritische Studie* (M. Niemeyer, Halle, 1906).
- Magoun, E. P. in *Early English and Norse Studies Presented to Hugh Smith in Honour of his Sixtieth Birthday* (eds Brown, A. & Foote, P.) 127–140 (Methuen, 1963).
- Kiernan, K. S. *Beowulf and the Beowulf Manuscript* (Rutgers Univ. Press, 1981).
- Liuzzi, R. M. in *Beowulf: Basic Readings* (ed. Baker, P.) 281–302 (Garland Publishing, 1995).
- Fitch, J. Sense-pauses and relative dating in Seneca, Sophocles and Shakespeare. *Am. J. Philol.* **102**, 289–307 (1981).
- Kuhn, H. Zur Wortstellung und -betonung im Altgermanischen. *Beitr. Gesch. Dtsch. Sprache Lit.* **57**, 1–109 (1933).
- Momma, H. The composition of Old English poetry. *Lang. Lit.* **7**, 175–178 (1998).
- Nagy, G. *Homer's Text and Language* (Univ. Illinois Press, 2004).
- West, M. L. *The Making of the Iliad: Disquisition and Analytical Commentary* (Oxford Univ. Press, 2011).

39. Sievers, E. *Altgermanische Metrik* (M. Niemeyer, Halle, 1893).
40. Gardner, T. The Old English kenning: a characteristic feature of Germanic poetical diction? *Mod. Philol.* **67**, 109–117 (1969).
41. Bosworth, J. *An Anglo-Saxon Dictionary* (Clarendon Press, 1989).
42. Cook, A. S. *The Christ of Cynewulf: A Poem in Three Parts: the Advent, the Ascension and the Last Judgment* (Ginn and Company, 1900).
43. Diamond, R. E. The diction of signed poems in Cynewulf. *Philolog. Q.* **38**, 228–241 (1959).
44. Schaar, C. *Critical Studies in the Cynewulf Group* (Haskell House, 1967).
45. Fulk, R. in *Cynewulf: Basic Readings* (ed. Bjork, R. E.) 3–22 (Garland Publishing, 1996).
46. Bjork, R. E. *The Old English Poems of Cynewulf* (Cambridge Univ. Press, 2013).
47. Stodnick, J. A. Cynewulf as author: medieval reality or modern myth? *Bull. J. Rylands Univ. Libr.* **79**, 25–39 (1997).
48. Carr, C. T. *Nominal Compounds in Germanic* (St. Andrews Univ., 1939).
49. Terasawa, J. *Nominal Compounds in Old English: A Metrical Approach* (Rosenkilde & Bagger, 1994).
50. Jockers, M. L. & Underwood, T. in *A New Companion to Digital Humanities* 2nd edn (eds Schreibman, S., Siemens, R. & Unsworth, J.) 291–306 (Wiley-Blackwell, 2016).
51. Greenberg, J. H. A quantitative approach to the morphological typology of language. *Int. J. Am. Ling.* **26**, 178–194 (1960).
52. O'Donoghue, H. *Old Norse-Icelandic Literature: A Short Introduction* (Wiley-Blackwell, 2004).
53. Bhrolchain, M. N. *An Introduction to Early Irish Literature* (Four Courts Press, 2017).
54. Zink, M. *Medieval French Literature: An Introduction* (Medieval and Renaissance Texts and Studies, 1995).
55. Dimock, W. C. Low epic. *Crit. Inq.* **39**, 614–631 (2013).
56. Russom, G. *Old English Meter and Linguistic Theory* (Cambridge Univ. Press, 1987).

Acknowledgements

The authors thank M. Nowak, S. Sinai and J. Gerold for helpful conversations, as well as S. Pintzuk and G. Russom for assistance in obtaining texts, dictionaries and scansions in formats amenable to computational analysis. This work was conducted under the auspices of the Quantitative Criticism Lab (www.qcrit.org), an interdisciplinary project co-directed by P.C. and J.P.D. and supported by a Neukom Institute for Computational Science CompX Grant and a National Endowment for the Humanities Digital Humanities Start-Up Grant (HD-248410-16). P.C. was supported by a New Directions Fellowship from the Andrew W. Mellon Foundation, and J.P.D. was supported by a National Science Foundation Graduate Research Fellowship (DGE1144152) and a Neukom Fellowship. The Program for Evolutionary Dynamics is supported in part by a gift from B. Wu and E. Larson. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

L.N., M.S.K., P.C. and J.P.D. designed the study. M.S.K., M.Y. and J.P.D. performed the study. All authors analysed the results. L.N., M.S.K., P.C. and J.P.D. wrote the manuscript, which was read and approved by all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41562-019-0570-1>.

Reprints and permissions information is available at www.nature.com/reprints.





Correspondence and requests for materials should be addressed to M.S.K. or J.P.D.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

In the format provided by the authors and unedited.

Large-scale quantitative profiling of the Old English verse tradition

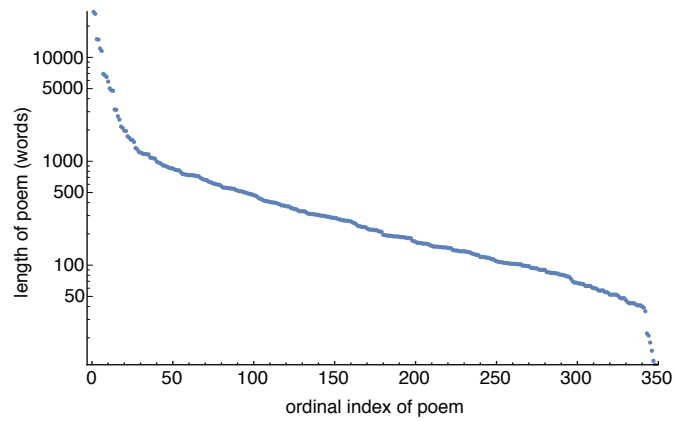
Leonard Neidorf ^{1,7}, Madison S. Krieger ^{2,7*}, Michelle Yakubek^{3,4}, Pramit Chaudhuri ⁵ and Joseph P. Dexter ^{6*}

¹Department of English, Nanjing University, Nanjing, China. ²Program for Evolutionary Dynamics, Harvard University, Cambridge, MA, USA.

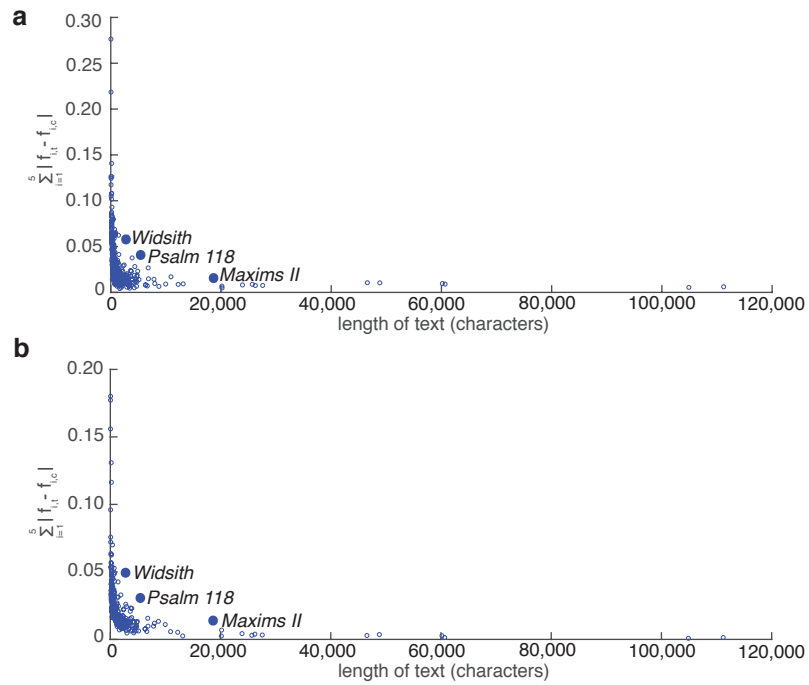
³Research Science Institute, Center for Excellence in Education, McLean, VA, USA. ⁴Texas Academy of Mathematics and Science, Denton, TX, USA.

⁵Department of Classics, University of Texas at Austin, Austin, TX, USA. ⁶Department of Systems Biology, Harvard Medical School, Boston, MA, USA.

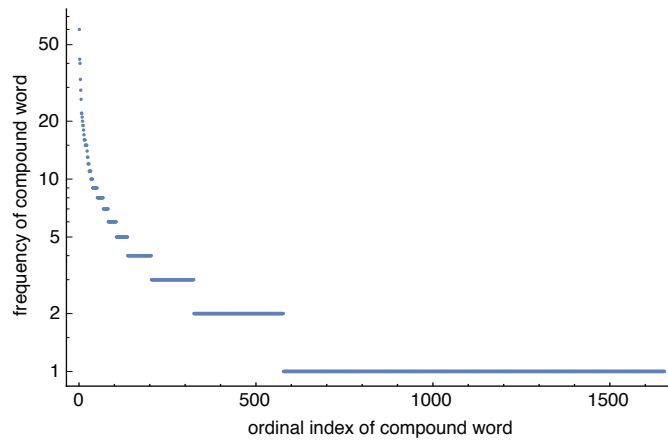
⁷These authors contributed equally: Leonard Neidorf, Madison S. Krieger. *e-mail: mkrieger@fas.harvard.edu; jdexter@fas.harvard.edu



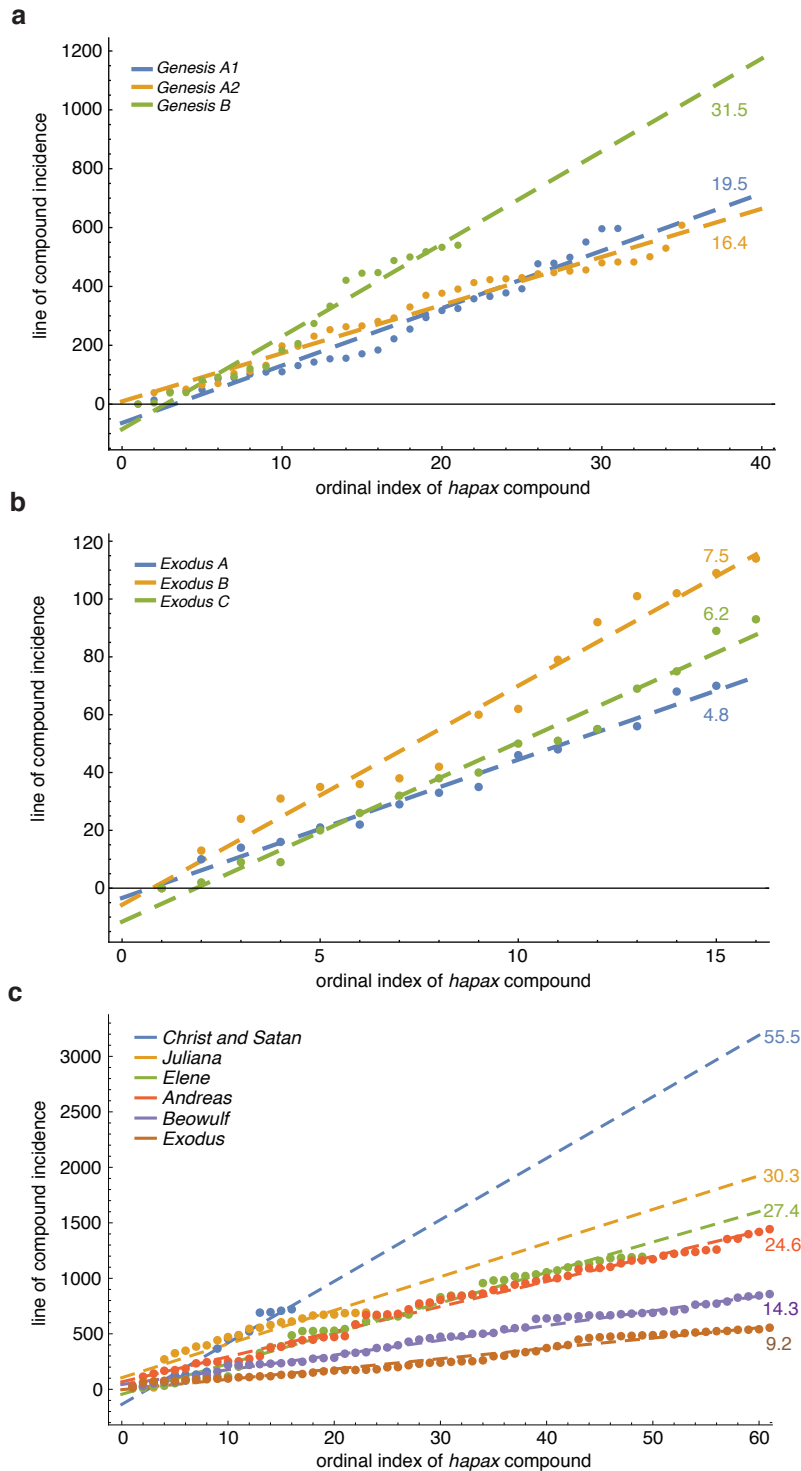
Supplementary Figure 1. Distribution of poems in the OE verse corpus by length. The OE verse corpus contains approximately 350 poems of total length 291,000 words. Aside from a small number of more substantial works, the vast majority of texts in the corpus contain fewer than 1,000 words.



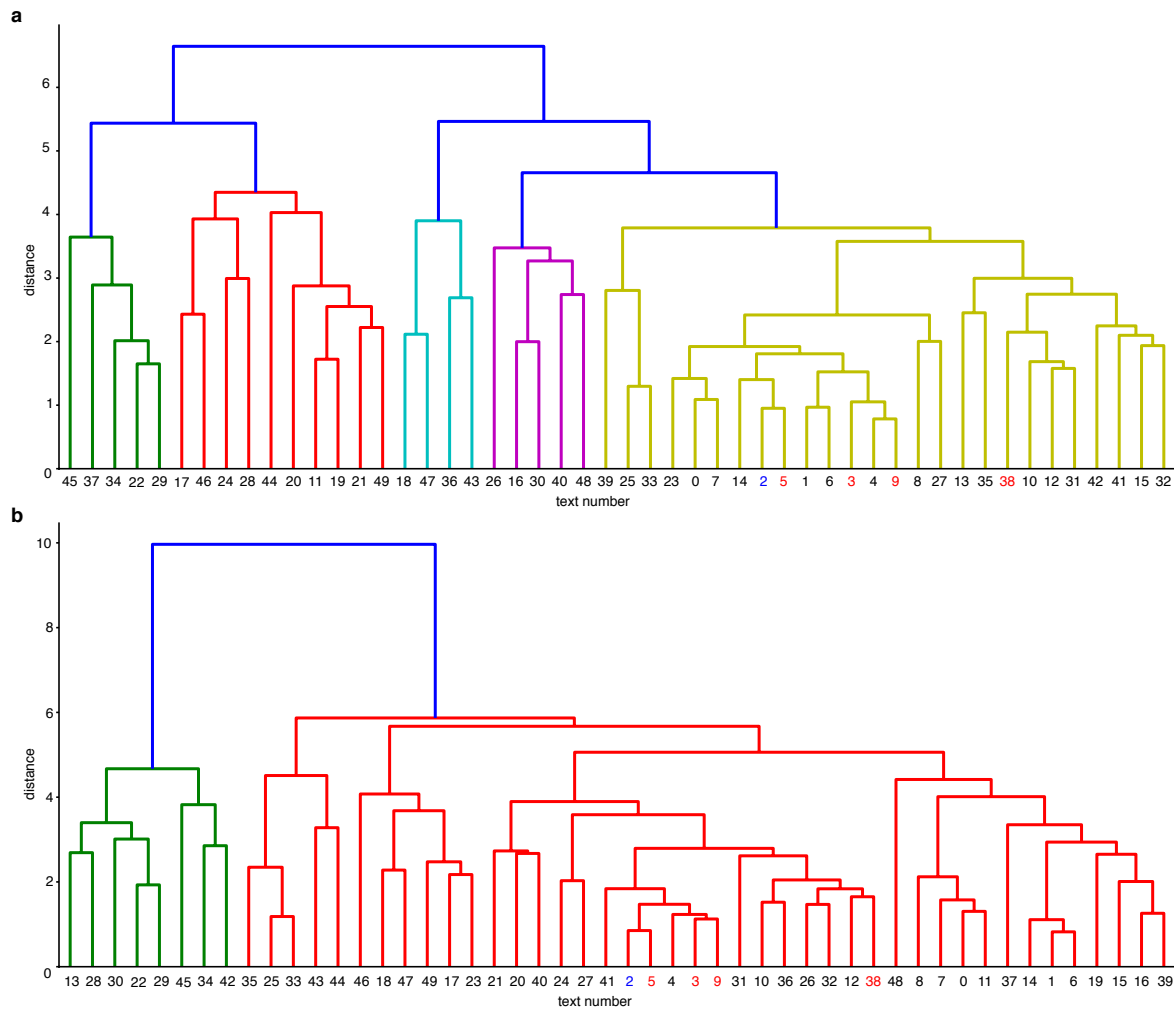
Supplementary Figure 2. Phonetic profiling using bigrams and four-grams. Plot of cumulative difference in functional n-gram frequency (for the five most common n-grams) against text length for **a** bigrams and **b** four-grams. As in Figure 1, each dot denotes one text, and anomalous texts are highlighted and labeled.



Supplementary Figure 3. Distribution of nominal compounds in the OE verse corpus. Most compounds are *hapax legomena* (bottom line).



Supplementary Figure 4. Usage of *hapax* compounds differs between authors. Rate of use of compounds in **a** three sections of the composite poem *Genesis* (Pearson's $r(36) = 0.973$, $r(32) = 0.981$, and $r(22) = 0.990$ for A1, A2, and B, respectively), **b** three random partitions of *Exodus*, which is believed to be of unitary authorship (Pearson's $r(10) = 0.995$, $r(10) = 0.971$, and $r(13) = 0.998$ for A, B, and C, respectively), and **c** a selection of longer poems, some written by Cynewulf (Pearson's $r(14) = 0.968$, $r(21) = 0.934$, $r(47) = 0.992$, $r(69) = 0.997$, $r(229) = 0.992$, and $r(62) = 0.990$ for *Christ and Satan*, *Juliana*, *Elene*, *Andreas*, *Beowulf*, and *Exodus*, respectively). $p < 0.001$ for all correlations by a two-tailed t-test. Numbers next to linear fits denote their slope.



Supplementary Figure 5. Additional dendrograms. Dendrograms produced from hierarchical agglomerative clustering with **a** functional bigrams and **b** functional four-grams. The numbering and color scheme for the texts is the same as in Figure 4 and corresponds to the labels in Supplementary Table 1.

Supplementary Table 1. List of texts as numbered in Figure 4 and Supplementary Figure 5.

Label	Poem
0	<i>Genesis</i>
1	<i>Beowulf</i> 1-2300
2	<i>Andreas</i>
3	<i>Christ</i>
4	<i>Guthlac</i>
5	<i>Elene</i>
6	<i>Beowulf</i> 2301-end
7	<i>Daniel</i>
8	<i>Christ and Satan</i>
9	<i>Juliana</i>
10	<i>The Phoenix</i>
11	<i>Exodus</i>
12	<i>Solomon and Saturn</i>
13	<i>Paris Psalm 118</i>
14	<i>Judith</i>
15	<i>The Battle of Maldon</i>
16	<i>The Judgment Day II</i>
17	<i>Meters of Boethius</i> 20
18	<i>Maxims I</i>
19	<i>The Menologium</i>
20	<i>The Seasons for Fasting</i>
21	<i>Azarias</i>
22	<i>Paris Psalm 77</i>
23	<i>The Dream of the Rood</i>
24	<i>Psalm 50</i>
25	<i>Soul and Body I</i>
26	<i>Widsith</i>
27	<i>The Descent into Hell</i>
28	<i>Paris Psalm 88</i>
29	<i>Paris Psalm 105</i>
30	<i>The Lord's Prayer II</i>
31	<i>The Judgment Day I</i>
32	<i>The Seafarer</i>
33	<i>Soul and Body II</i>
34	<i>Paris Psalm 106</i>
35	<i>Resignation</i>
36	<i>The Wanderer</i>
37	<i>Paris Psalm 104</i>
38	<i>Fates of the Apostles</i>
39	<i>Meters of Boethius</i> 26
40	<i>The Gifts of Men</i>
41	<i>The Order of the World</i>
42	<i>Paris Psalm 68</i>
43	<i>Riddle 40</i>
44	<i>Metrical Charm I</i>
45	<i>Paris Psalm 103</i>
46	<i>Meters of Boethius</i> 11
47	<i>The Fortunes of Men</i>
48	<i>The Rune Poem</i>
49	<i>Meters of Boethius</i> 29

Nature Research, brought to you courtesy of Springer Nature Limited (“Nature Research”)

Terms and Conditions

Nature Research supports a reasonable amount of sharing of content by authors, subscribers and authorised or authenticated users (“Users”), for small-scale personal, non-commercial use provided that you respect and maintain all copyright, trade and service marks and other proprietary notices. By accessing, viewing or using the nature content you agree to these terms of use (“Terms”). For these purposes, Nature Research considers academic use (by researchers and students) to be non-commercial.

These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). By sharing, or receiving the content from a shared source, Users agree to be bound by these Terms.

We collect and use personal data to provide access to the nature content. ResearchGate may also use these personal data internally within ResearchGate and share it with Nature Research, in an anonymised way, for purposes of tracking, analysis and reporting. Nature Research will not otherwise disclose your personal data unless we have your permission as detailed in the Privacy Policy.

Users and the recipients of the nature content may not:

1. use the nature content for the purpose of providing other users with access to content on a regular or large scale basis or as a means to circumvent access control;
2. use the nature content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval, sponsorship, or association unless explicitly agreed to by either Nature Research or ResearchGate in writing;
4. use bots or other automated methods to access the nature content or redirect messages; or
5. override any security feature or exclusionary protocol.

These terms of use are reviewed regularly and may be amended at any time. We are not obligated to publish any information or content and may remove it or features or functionality at our sole discretion, at any time with or without notice. We may revoke this licence to you at any time and remove access to any copies of the shared content which have been saved.

Sharing of the nature content may not be done in order to create substitute for our own products or services or a systematic database of our content. Furthermore, we do not allow the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Nature content cannot be used for inter-library loans and librarians may not upload nature content on a large scale into their, or any other, institutional repository.

To the fullest extent permitted by law Nature Research makes no warranties, representations or guarantees to Users, either express or implied with respect to the nature content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.

Please note that these rights do not automatically extend to content, data or other material published by Nature Research that we license from third parties.

If you intend to distribute our content to a wider audience on a regular basis or in any other manner not expressly permitted by these Terms please contact us at

onlineservice@springernature.com

The Nature trademark is a registered trademark of Springer Nature Limited.

Across America, artists are searching for answers about Trump's planned funding cuts

President proposed slashing funds for arts, humanities endowments, Corporation for Public Broadcasting

By Haydn Watters, [CBC News](#) Posted: Mar 26, 2017 11:13 AM ET Last Updated: Mar 26, 2017 12:29 PM ET



There's no time to wait-and-see. American artists, researchers and media people are already planning what comes next should President Donald Trump's plan to cut all funding for the National Endowment for the Arts (NEA), National Endowment for the Humanities (NEH) and the Corporation for Public Broadcasting (CPB) pass through Congress.

The president [laid out the plan in his proposed budget](#) earlier this month. The funds the trio of organizations gets from Washington make up a miniscule portion of the overall federal budget — in the 2016 fiscal year total federal spending was [an estimated \\$3.9 trillion US](#). The NEA got \$147.9 million US and [the NEH requested the same amount](#) (that's about 0.004 per cent each) while the CPB received \$445 million US (around 0.01 per cent).

That's not a lot of money next to the trillions spent overall, but it's vital for hundreds of recipients, scattered in galleries, universities, museums, radio booths and television studios throughout the United States.

CBC News spoke to five different groups — pictured on the map above — who were recent recipients of funding from the NEA, NEH and the CPB, about what the money meant to them and what would happen if it was pulled.

Jackson Hole Public Art (Jackson, Wyo.)

Most recent NEA grant: \$50,000 to plan for public art in a part of town that's being redesigned. At the moment, the project is a wildlife viewing platform, designed by sculptor Buster Simpson.

What the money means: Carrie Geraci, the program's director, said the funding has brought in big money from other donors. "I think when anyone receives an NEA grant, anyone in the arts world and anyone who invests in projects like this understand the high level of scrutiny they go through."

Geraci said the grants are "vital" in small towns like Jackson, Wyo., home to about 10,000 people and a lot of artists. She said the high cost of living can make it hard for artists to make a living.

What would happen if it got pulled? "We wouldn't do this project. We wouldn't have enough funds," she said.

Geraci said any cut would have a bigger impact on smaller towns like hers, which don't have as easy

access to other funders as urban centres.

Her message for Trump: "It's just absolutely unbelievably shortsighted to think that this type of investment does not have an economic impact. It's felt just as strongly in rural communities as it is in urban ones," she said. "I would tell him to look at his own children and [see] how they have benefited from arts and culture education."

Ideastream (Cleveland)

Most recent CPB grant: About \$2 million a year, divvied up between the three public media stations it runs — PBS member WVIZ, NPR member WCPN and classical station WCLV.

What the money means: Kevin Martin, Ideastream's president and CEO, said it makes up eight per cent of the annual operating budget and goes into programming.

"There is just no viable alternative to replace those dollars," he said, adding that it is a "mystery" why Trump wants to cut funding.

How he's feeling: Martin's faced cuts before, but said he is particularly worried this time because of the administration's "unpredictability."

He remains optimistic: "We have broad bipartisan support and sometimes the president's budget is used as a statement."

His message for Trump: "Public media is a national treasure. I think citizens all over the country rely on this service ... I don't know of another service or agency that the government funds that yields that kind of return."

Quantitative Criticism Lab (Austin, Texas)

Most recent NEH grant: \$74,921 to fund the lab's research into parallels between computational biology and classical literature. The lab is researching literature using scientific techniques from biology, hoping to find out more about the ancient texts and be able to share that with others.

What the money means: Pramit Chaudhuri, who co-directs the lab with Joseph Dexter at the University of Texas at Austin, said the money goes towards recruiting and paying people with the broad set of skills the project needs.

"If funding like this isn't available, I won't say it's impossible, but it certainly raises the bar," said Chaudhuri.

What would happen if it got pulled? "It could be devastating for projects like mine," he said. "I think after a very small amount of time, we would no longer be able to keep doing it."

And though there's no guarantee the cuts will happen, Chaudhuri's already been looking into other

funding models, like individual donations.

His message for Trump: "He certainly talks a great deal about the strengths of the United States, making the case for a kind of U.S. exceptionalism. And regardless of what you think about that, one of the ways the U.S. has been able to do that is through education."

A Studio in the Woods (New Orleans)

Most recent NEA grant: The studio just got two — \$50,000 for a neighbourhood fruit tree planting project and \$15,000 for artist residencies.

What the money means: Ama Rogan, the director of the artist community which sits on the Mississippi River, said the grant helps validate what they do. "Those monies are really key to us because we directly fund artists, we put money into their hands," she said.

"It means something to us to have something that's national funding. It's like, we're representing the country's interest in what we are doing."

How she's feeling: "The conversations are still in a 'Oh my god, I can't believe this is happening' mode," she said. "We're not giving up on the NEA."

Rogan's prepared to "hustle" to keep it going.

Her message for Trump: "I think the NEA should continue to be funded because arts are an integral part of this country and the citizens within it. Art is the way we are able tell our stories to each other."

Catticus Corp. (Berkeley, Calif.)

Most recent NEH grant: \$400,000 to make *Mad as Hell!*, a documentary about the 1978 California tax revolt.

What the money means: Jason Cohn, the director, said his career as a documentary filmmaker has depended on NEH funding — he thinks this is the fifth or sixth film he has worked on that's received funding and he's been rejected many other times.

"I don't think there's any way to make this film without the NEH funding," he said. "There's this idea that it's an elite thing and it's just so misguided and wrong ... it's a ridiculous place to look for savings."

What would happen if it got pulled? "I personally will have to come up with a new way of making films," he said. Cohn has given it some thought; he said if NEH funding is killed entirely, it would also kill history documentary filmmaking in the U.S.

"At the moment, these films simply don't get made without the endowment."

His message for Trump: "I would just say that culture is crucial to a civilization," he said. "Our history matters and our culture that we create matters and if we don't have ways of sharing it and making it

accessible to everyone, then something that's just incomprehensibly valuable is lost."

- [Winners and losers in Trump's budget](#)
- [Alaska public broadcasters fear funding cuts under Trump administration](#)

Explore CBC

CBC Home

TV

Radio

News

Sports

Music

Life

Arts

Kids

Local

Documentaries

Comedy

Books

Parents

Indigenous

Digital Archives

Games

Contests

Site Map

Stay Connected

Apps

RSS

Podcasts

Newsletters & Alerts

Services and Information

Corporate Info

Public Appearances

Commercial Services

Reuse & Permission

Terms of Use

[Privacy Policy](#)

[CBC Shop](#)

[Help](#)

[Contact Us](#)

[Jobs](#)

[Doing Business with Us](#)

[Renting Facilities](#)

CBC  **Radio-Canada**

©2017 CBC/Radio-Canada. All rights reserved

[Visitez Radio-Canada.ca](#)

Harvard Medicine



Joseph Dexter

A Closer Read

To understand large data sets, researchers look to tools that decipher patterns in natural language

by Kevin Jiang

On a miserably cold January evening in 2014, Joseph Dexter met his friend and mentor Prमित Chaudhuri at a party at a classical studies conference in Chicago. Dexter, a graduate student in the HMS Department of Systems Biology, had met the former

Dartmouth classics professor when Dexter, while still in high school, was taking classes at the New Hampshire college. Catching up with one another that evening began with the usual pleasantries, but their conversation soon carried them into uncharted territory. As the night deepened and the room emptied, the pair remained huddled, deliberating an idea: could bioinformatics be adapted for studying ancient literature?

In the first century CE, the Roman philosopher and statesman Seneca—tutor, advisor, and, ultimately, victim of Emperor Nero’s anger—wrote a series of plays shaped by the political and social strife of his era. Collectively known as the Senecan tragedies, this corpus was relegated to the margins of history until it was rediscovered by Renaissance scholars in the fifteenth century. The plays’ reemergence marked the revival of the tragedy on European stages and served as a model for dramatic traditions that influence Western culture to this day.

The journey of the Senecan tragedies from antiquity to modernity has taken unpredictable turns. But perhaps the unlikeliest detour was made during that late-night conversation, where, over several glasses of wine, a biologist and a classics scholar began to flesh out how techniques from bioinformatics could be used to gain insights about texts such as the Senecan tragedies.

At first blush, it might seem implausible to speculate that ancient Roman plays packed with supernatural intervention and bloodthirsty revenge would have anything in common with the computational analysis of biological data. But for Dexter, whose lifelong obsession with classics paralleled his path to the study of math and biology, and for an increasing number of researchers like him, the intersection of computation, human language, and biology is fertile ground for discovery.

“There are lots of commonalities that arise when you deal with large amounts of multidimensional data in messy, unstructured contexts,” he says. “That’s certainly true in biomedicine, and it’s certainly true in culture and literature.”

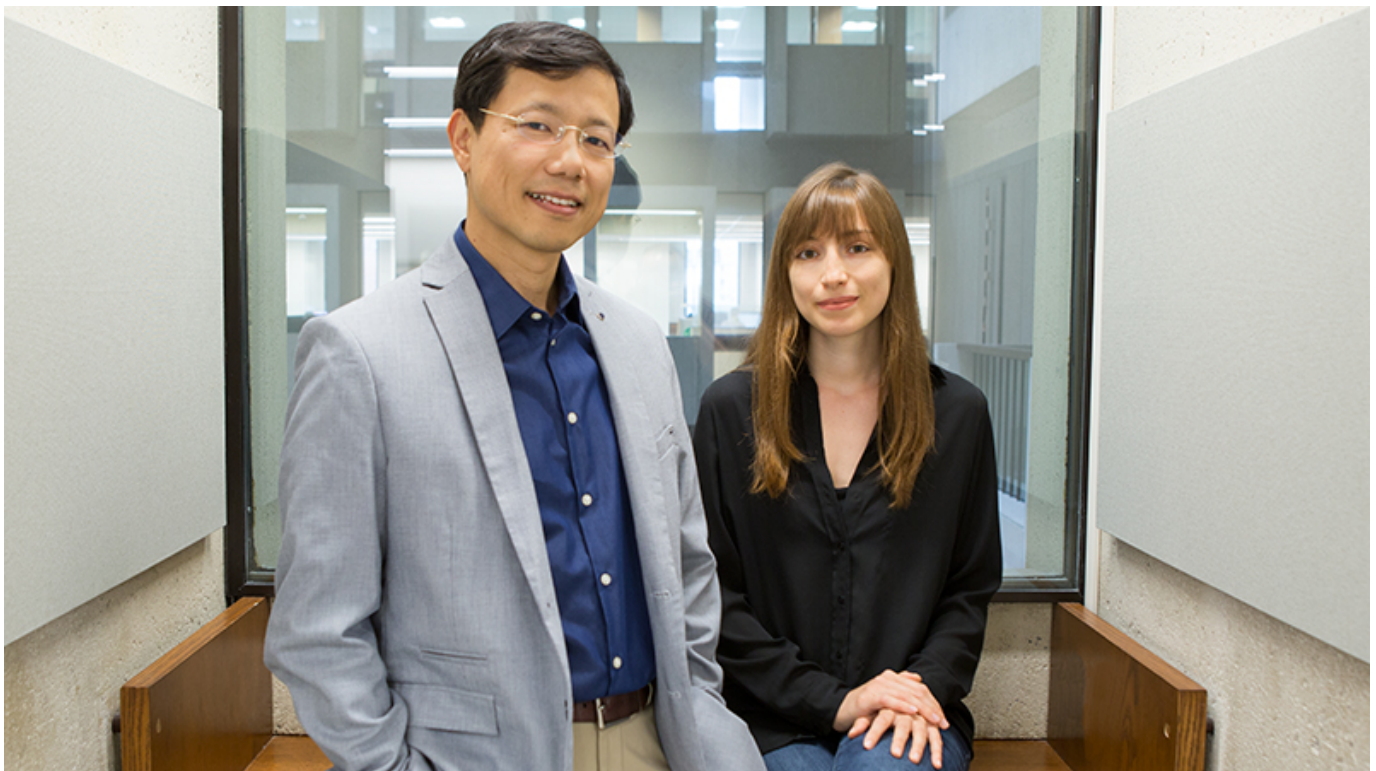
Driven by rapid growth in computing power and new technologies, almost every facet of biomedical research has been deluged with data in recent years, from the petabyte-sized datasets of “-omic” fields used to study the genome, transcriptome, proteome, and similar molecular entities, to what many are estimating will become the zettabyte-sized data sets of scientific literature and electronic medical records (EMRs).

Extracting meaningful discoveries out of this wealth of information has necessitated the development of tools that not only can identify patterns of interest across massive data sets but can do so despite the inherent “messiness” of biology. This is no simple challenge. Whether at the level of molecules or populations, the study of biological systems involves untangling sets of rules, connections, and interdependencies that have been laid down by evolution that can vary by timing, context, and chance.

Yet computational techniques honed for the study of the complex, interconnected, often ambiguous system we call language are increasingly being used to inform biomedical research. For some applications, these tools are showing enormous promise, from improving our understanding of genomics and biochemical pathways to realizing the full potential of precision medicine.

Dramatic Language

As early as the 1940s, linguists and computer scientists were collaborating on methods that would allow computers to learn, understand, and apply human language to a variety of uses. Known as natural language processing, researchers drew from disciplines such as artificial intelligence, machine learning, computer science, statistics, and computational linguistics to analyze the rules and patterns of language.



Peter Park and Doga Gulhan

As large amounts of linguistic data and increased computing power became available, these efforts bloomed, leading to contemporary applications such as Siri, Apple's intelligent personal assistant software, and Google Translate. The field of biomedical informatics, which leverages similar techniques to analyze and interpret medical and biological data, has similarly matured over the past few decades.

Natural language processing and biomedical informatics intersect in many ways. One of the more unusual examples may be the project launched by Dexter and Chaudhuri after that late-night conversation. Applying a technique they dubbed quantitative literary criticism, the project's team of classics scholars, computer scientists, and computational biologists used computational tools to analyze ancient Latin and Greek texts, including the plays by Seneca.

Earlier this year, Dexter and his colleagues published a paper in *Proceedings of the National Academy of Sciences* in which they used computational profiling of writing style to explore intertextuality—the concept that all texts have relationships to other texts—across the writings of ancient authors. In one trial, they computationally analyzed the entirety of the Senecan tragedies to investigate their influence on a play by a fifteenth-century Italian author writing in the Senecan tradition. The team identified places in which the later play differs in style from plays written by Seneca. By pinpointing these differences, they could reveal various literary effects for which the author was striving and which gave his work its distinctive character.

The group is also pursuing a method for the detection of verbal intertextuality based on one of the most common bioinformatics techniques: sequence alignment. This analysis allows like-to-like comparisons of DNA, RNA, or protein sequences by lining up the molecular strands so that they match at as many locations as possible. In evolutionary studies, this technique has been used to identify similar genes across different species and analyze the degree of difference between them to build phylogenetic trees.

“Linguistics played an important role in the development of sequence alignment tools that are now ubiquitous in biology” says Dexter. “We realized you could use the same techniques on literary problems.”

Topic Sentences

Bioinformatics tools can have powerful and creative applications, but when combined with natural language processing and applied to biomedical sciences, they have profound implications for human health.

On the third floor of the Francis A. Countway Library of Medicine, Peter Park, an HMS professor in the Department of Biomedical Informatics, oversees a research group that is using large-scale computational analysis of genomics data to better understand the mechanisms underlying human diseases.

Among the group's many approaches is one drawn directly from natural language processing: a statistical model that can identify what "topics" are contained within texts. Instead of analyzing language, however, Park and his team are identifying the specific causes of mutations in the genomes of cancer patients.

To illustrate with an analogy, a book about military battles of World War I will include the words "tank" and "trench" more frequently than a book about battles in the American Revolutionary War. But both will have more occurrences of words like "gun" and "cannon" than a book about the Punic Wars, which raged in the third through second century BCE.

This technique can be used to scan entire libraries of literary texts for groups of co-occurring words that indicate a common topic. The statistics can then be used to infer not only what the topic of a book may be, but the mixture of topics contained within.

Led by HMS bioinformatics postdoctoral fellow Doga Gulhan—a particle physicist who trained at MIT and worked at CERN—the team applied this concept to genomes. Key to their work are studies that have linked certain causal factors to specific patterns of mutations. In the genomes of smokers, for instance, there is a dramatic increase in cytosine to adenine mutations. These single nucleotide variants are often accompanied by predictable patterns in nucleotides on either side of the single variant.

"If we think of each person's genome as a book that contains many mutations or words," says Gulhan, "we can use our algorithms to find words that occur together and group them by common occurrences into broad topics. You cannot do this using only a few genomes. You need a big set of books so that you can determine what the topics are. Then you can look at each genome to see which topics it contains."

Park, Gulhan, and their team are scanning trillions of DNA base pairs and petabytes of data found in roughly 2,700 different tumor genome sequences from the International Cancer Genome Consortium. They have identified dozens of mutation signatures that indicate different causal factors, or "topics," in their analogy. Most of these factors are still unknown, but some, including smoking and UV exposure, have been previously identified and are being used to validate and improve the methodology.

“Ultimately, what we want to do is give patients treatments that are appropriate for their disease,” Park says. “If you are presented with two tumors, say, a brain tumor and a lung tumor, they might appear to be caused by different factors. But it could be that the same mechanism is causing mutations in both. Sequencing the genomics of cancer patients will soon be a routine practice, and this type of genome analysis will help us sift through the mutations that reflect the history of the tumor, so that we can identify the best drug or combination of drugs to use for the patient.”

WALL-E

The tools of natural language processing have shown great promise when applied to biological data, but they are no less valuable within the context of their original intent: to provide computers with the capability to do useful things with human language.

Since 2005, the number of papers and abstracts on biomedical topics indexed by the National Institutes of Health’s PubMed search engine has doubled, sitting at somewhere around 27 million, with thousands more being added daily.

“Scientific literature is growing so large that we can’t keep up with it all, even within fields,” says John Bachman, a research fellow in therapeutic science in the Laboratory for Systems Pharmacology (LSP) and the Harvard Program in Therapeutic Science (HiTS) at HMS. “And it is extremely difficult to know if something relevant to your research might exist in some other field.”

In 2014, DARPA, a research and development wing of the U.S. Department of Defense, launched a project to address this growing concern. Dubbed the Big Mechanism program, DARPA tasked research teams with developing computational tools that could intelligently scan and make sense of scientific literature.

To tackle this challenge, a group led by Peter Sorger, the Otto Kraymer Professor of Systems Pharmacology at HMS and director of the LSP and HiTS, relied heavily on natural language processing. Led by Bachman and Benjamin Gyori, a research fellow in therapeutic science in the LSP, the team is developing a software platform that reads papers and builds models of complex biochemical networks and can also support interactive dialog with scientists in a manner akin to Apple’s Siri.



John Bachman (left) and Benjamin Gyori

The platform, named INDRA (the Integrated Network and Dynamical Reasoning Assembler) first uses machine language to parse scientific publications and abstracts to look for phrases of interest. These phrases can include biochemical names and processes, as well as key words, for example, “tumorigenesis” or “metastasis.”

“When these systems extract information from the literature, it comes out as this big, error-prone, redundant, fragmented bag of facts,” Gyori says. “The main goal of INDRA is to turn those facts into coherent, predictive, and explanatory models. We’re not just looking for statistical associations in text, like co-occurrence of a drug name with a disease name. We want to extract causal events.”

To do so, the team developed what they’re calling a knowledge assembly methodology. INDRA cross-references raw phrases against each other as well as against databases and other knowledge sources in a manner analogous to sequence alignment. Guided

by sophisticated algorithms, INDRA eliminates redundant statements and likely errors about biological processes and identifies the mechanisms that connect them.

The scale at which INDRA can do this is difficult, if not impossible, for humans to achieve. In one proof-of-concept trial, INDRA assembled a biochemical network model after scanning a corpus of 95,000 papers that contained information relevant to a single study of interest. This study reported on tests involving the efficacy of nearly one hundred drug combinations on melanoma cell lines from which the twenty-two strongest drug effects were selected. The team asked INDRA to find the mechanisms involved. Of the twenty-two observed effects of a drug on a protein, INDRA generated detailed biochemical explanations for twenty, a 90 percent success rate.

With additional natural language processing development, the team has devised a software prototype, provisionally named Bob, that one day will allow any scientist to ask INDRA questions in English and receive an answer in English, basically a virtual lab assistant that can supply information to help researchers formulate and evaluate hypotheses.

Syntax

For patients, tools like INDRA and the topic model used by Park and Gulhan have tremendous potential in opening new lines of research and discovery that can someday affect their health and quality of life. But natural language processing can also have a direct benefit at the bedside.

Perhaps the largest data sets that exist in the biomedical sciences are EMRs, which contain clinical narratives and details such as disease pathology and treatments for hundreds of millions of patients. There is, however, no universal system for EMRs, so they can differ greatly in how critical data elements are presented, from coding for medications to vocabulary use.

This lack of conformity presents an ideal problem for natural language processing tools, one that Guergana Savova, an HMS associate professor and director of the Natural Language Processing Lab at Boston Children's Hospital, may help solve. Savova and her colleagues are building systems that can read and analyze anonymized clinical notes from EMRs and combine that information with other types of information.

One of their efforts is aimed at performing “deep phenotyping” on cancer. Through their analysis of the plain text within millions of EMRs, they hope to reveal the relationships between the characteristics of a cancer, including its molecular profile, grade, and metastasis patterns, and information extracted about patients, such as family histories, tests, treatments, and comorbidities.

“We need to learn as much as we can about these connections if we are to achieve the goal of precision medicine, because every patient and every tumor has a different set of characteristics,” says Savova, a computational linguist and computer scientist by training. “These questions can be answered only if researchers have large corpora of data from large cohorts of patients to compare. Manually, it’s just not doable.”



Alexa McCray

But state-of-the-art natural language processing systems are not a panacea, and no system is perfect, Savova says. Although errors can be controlled for—INDRA, for example, has a “belief engine” to allow it to determine its probability of correctness—inaccuracies arise for a variety of reasons that range from language variations to the differences in statistical and computational algorithms that underlie any given system.

“We build extraction tools, but there is a tremendous difference between extraction of information and such a complex decision-making process as diagnosis,” Savova says. “What a physician observes or hears or feels, the logical and creative steps that humans are capable of, are not necessarily recorded in the EMRs, and they are as important as any amount of text processing. The big question for artificial intelligence in general is how to encode this comprehensive knowledge into one representation.”

The vast majority of current-generation natural language processing systems rely on human-initiated resources, such as a list of Latin phrases or biochemical names to search for in a corpus of data or a backbone of medical terms to which clinical notes are connected. This can be a troubling variable.

“There are people who disagree with me,” says Alexa McCray, a professor of medicine in the Department of Biomedical Informatics at HMS and Beth Israel Deaconess Medical Center, “but if you’re working with not-so-good data on the way in, then what comes out the other end is not going to be so good either.”

Ensuring access to high-quality data for computational applications has been a priority for McCray for almost her entire career. A linguist who joined IBM as the field of computational linguistics was blossoming, McCray spent decades at the National Library of Medicine at the NIH.

There, she helped develop standards such as the Unified Medical Language System, a comprehensive and curated database of millions of biomedical concepts and names. That system now serves as the backbone for many natural language processing applications.

For biomedical researchers to make full use of natural language processing and uncover knowledge that can affect human health and disease, there must be a strong foundation of data built through human effort.

“Data standards, curation, and language processing, these are areas where I think we have to put more of our combined energy,” McCray says. “Otherwise, it’s the Tower of Babel. What we need to get to is a point where we can compare apples to apples across biomedicine.”

Kevin Jiang is a science writer in the HMS Office of Communications and External Relations.

Images: John Soares

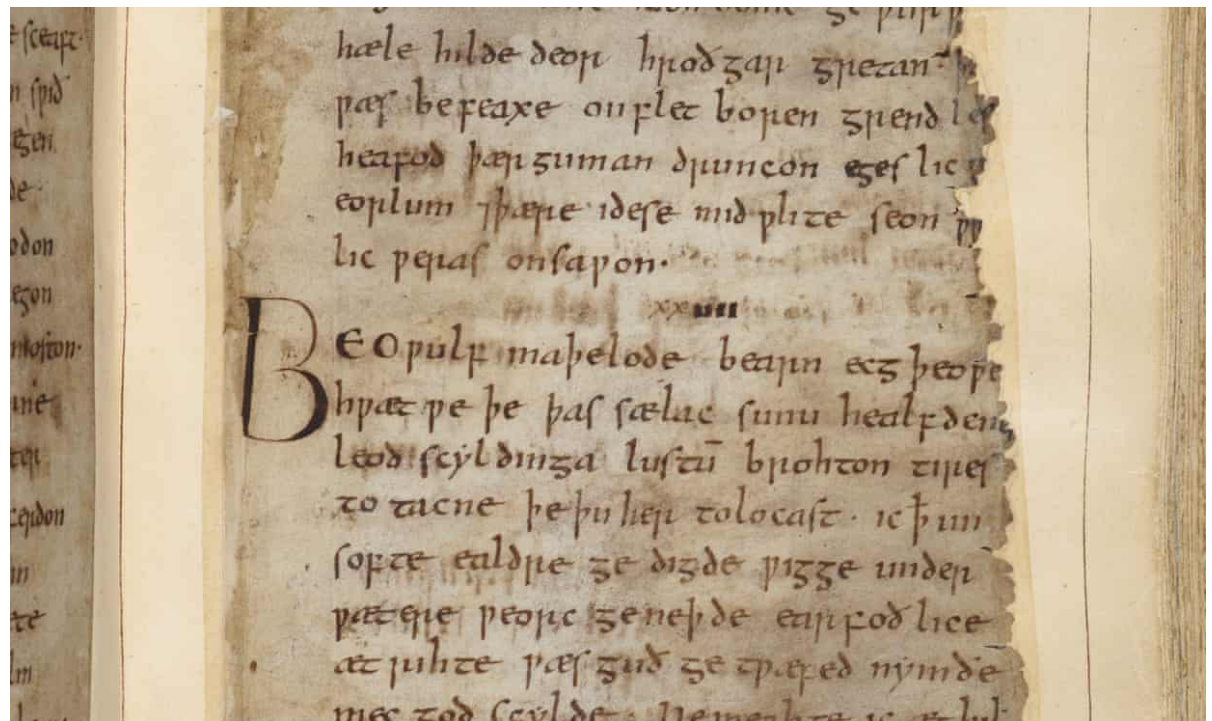
The Guardian

Beowulf the work of single author, research suggests

Debate over whether poem was written by multiple authors or one has raged for years

Nicola Davis

Mon 8 Apr 2019
11.00 EDT



The Beowulf manuscript at the British Library in London. Photograph: British Library

Beowulf, the epic poem of derring-do and monsters, was composed by a single author, research suggests, pouring cold water on the idea it was stitched together from two poems.

One of the most famous works in Old English, Beowulf tells of the eponymous hero who defeats the monster Grendel and his mother, thereby rescuing the Danes from a reign of terror, before returning to his homeland and later dying in a battle with a dragon.

But the poem has been the subject of a long-running debate. While some argued the work is the product of multiple poets, others - including the scholar and Lord of the Rings author JRR Tolkien - have said the evidence suggests it is a single poet's work.

Recently the debate has resurfaced with some suggesting the poem is the result of two

different works joined together – one involving Beowulf’s escapades in Denmark and one involving the dragon.

Now a study adds to a growing body of work suggesting Beowulf was composed by just one poet.

“The authorship question is a topic of perennial interest in Beowulf studies,” said Leonard Neidorf, professor of English Literature at Nanjing University and co-author of the research. “Our article reopens the question in order to apply for the first time some of the most sophisticated computational methods available for author identification.”

Writing in the journal *Nature Human Behaviour*, the team of researchers from the US and China report how they came to their conclusion by splitting the poem into the two pieces around the points where scholars have suggested a split, and analysing small features of the text.

While various aspects of the poem, including word use, themes and style, have been explored before, the latest study looks at even smaller features of the text and their patterns of use. These include the use of certain types of pause, the use of different rhythms, and the occurrence of words produced by joining others together – such as “bone-house” (written as *ban-hus*), which the authors say was used to mean the human body. The team also looked at the use of clusters of letters found within words, which are important for the sound of a poem.

The results, derived from computer-based analysis, reveal striking similarities in the way such features were used across both sections of the text. That suggests – although cannot conclusively prove – it was the work of a single poet, the researchers say.

By contrast, the Old English epic *Genesis*, which is believed to be the product of more than one poet, was found to have marked differences, both in terms of the patterns of the pauses and the use of compound words, between what are thought to be its constituent parts.

But mysteries remain – not least the identity of the Beowulf author. “The most that can be inferred from the language of the poem is that the author probably spoke the Mercian dialect and probably lived during the first half of the eighth century,” said Madison Krieger, co-author of the study from Harvard University.

As well as the findings about Beowulf, the team says the approach also supports the controversial claim that the Old English poem *Andreas*, which charts the dramatic exploits of St Andrew, was composed by a poet called Cynewulf, who is believed to have created at least four other works based on religion.

“With Cynewulf, our tests encourage scholars to reconsider a possibility that has not been seriously entertained in the past half century,” the researchers write.

Dr Francis Leneghan, a Beowulf expert at the University of Oxford, said the study joined a body

of evidence supporting the view that Beowulf was composed by one poet. However, he said it would be useful to apply the analysis to smaller chunks of the text to test the idea that it might have been formed from many smaller poems stitched together, or that some lines might have been added over the centuries by scribes.

Leneghan said the authors' conclusions around Andreas were less convincing, and would stir debate, noting it was thought that the author of Andreas had almost certainly read Beowulf and the works of Cynewulf. "Resemblances between Andreas and the works of Cynewulf are more likely to be the result of imitation," he said.

Kruger stressed the results were not definitive. "We absolutely entertain the idea that Andreas could be written by a Cynewulf imitator," he said. "Our work just suggests this might be a less likely explanation than scholars have believed in the past."

Since you're here...

... we have a small favour to ask. More people are reading and supporting our independent, investigative reporting than ever before. And unlike many news organisations, we have chosen an approach that allows us to keep our journalism accessible to all, regardless of where they live or what they can afford.

The Guardian is editorially independent, meaning we set our own agenda. Our journalism is free from commercial bias and not influenced by billionaire owners, politicians or shareholders. No one edits our editor. No one steers our opinion. This is important as it enables us to give a voice to those less heard, challenge the powerful and hold them to account. It's what makes us different to so many others in the media, at a time when factual, honest reporting is critical.

Every contribution we receive from readers like you, big or small, goes directly into funding our journalism. This support enables us to keep working as we do - but we must maintain and build on it for every year to come. **Support The Guardian from as little as \$1 - and it only takes a minute. Thank you.**

Support The Guardian



Topics

- Poetry
- news

MII

122

Final

Next Score

'Beowulf' is bloody, canonical, and long — and one person wrote it, scholars say

By [Travis Andersen](#) Globe Staff, April 11, 2019, 12:26 p.m.



Grendel, the monster, is vanquished by the hero, Beowulf, in the Old English classic. (JOHN HENRY FREDERICK BACON)

Only one person created the monster.

That's according to a team of researchers at Harvard, Dartmouth, and elsewhere, who determined the epic poem ["Beowulf,"](#) a staple of literature classes the world over, was written by a sole author more than a millennium ago.

The findings of the team, led by Madison Krieger, a postdoctoral fellow at Harvard's Program for

Evolutionary Dynamics, and Joseph Dexter, a Harvard PhD who's now a Neukom fellow at Dartmouth College, were published April 8 in [the journal Nature Human Behaviour](#), Harvard said in a statement.

While the poem itself, which features a bloody clash between the hero, Beowulf, and Grendel, a mythical monster, is wicked old, the researchers arrived at their findings with the aid of a computer and cutting-edge algorithms.

“Using a statistical approach known as stylometry, which analyzes everything from the poem’s meter to the number of times various combinations of letters show up in the text, Krieger and his colleagues found new evidence that ‘Beowulf’ is the work of a single author,” the university said.

Krieger said the research team conducted a meticulous review of the text.

“We looked at four broad categories of items in the text,” he said in the statement. “Each line has a meter, and many lines have what we call a sense pause, which is a small pause between clauses and sentences similar to the pauses we typically mark with punctuation in modern English. We also looked at aspects of word choice.”

He continued, “But it turns out one of the best markers you can measure is not at the level of words, but at the level of letter combinations. So we counted all the times the author used the combination ‘ab,’ ‘ac,’ ‘ad,’ and so on.”

Krieger added that across “many of the proposed breaks in the poem, we see that these measures are homogeneous. So as far as the actual text of Beowulf is concerned, it doesn’t act as though there is supposed to be a major stylistic change at these breaks. The absence of major stylistic

shifts is an argument for unity.”

Questions surrounding authorship of “Beowulf” have long divided academics, and the debate is expected to continue, the study’s findings notwithstanding.

“If we really believe this is one coherent work by one person, what does it mean that it has these strange asides?” Krieger said. “Maybe one of the biggest takeaways from this is about how you structured a story back then. Maybe we have just lost the ability to read literature in the way people at the time would have understood it, and we should try to understand how these asides actually fit into the story.”

The study also credits Leonard Neidorf, an English professor at Nanjing University; Michelle Yakubek, who contributed as a student at MIT’s Research Science Institute; and Pramit Chaudhuri, associate professor of classics at the University of Texas at Austin.

Chaudhuri and Dexter are the codirectors of the Quantitative Criticism Lab, “a multi-institutional group devoted to developing computational approaches for the study of literature and culture,” the university said.

Right now, you might be reaching deep into the recesses of your brain to recall highlights of the grim narrative poem, as framed by your high school English teacher.

Here’s a primer:

Written more than 1,000 years ago in Old English, “Beowulf” recounts the deeds of the warrior Beowulf, who goes to the aid of the Danes after they’re terrorized by a monster called Grendel. When Beowulf slays Grendel, he’s embraced by King Hrothgar, who looks on him like a son.

When Grendel's mother arrives to avenge her son's death, Beowulf kills her, too.

He returns to his home (in present-day Sweden), where he rules his people for 50 years until the lair of a dragon is disturbed. This time, Beowulf receives a mortal blow and although he kills the dragon, he, too, dies.

But we leave you, gentle reader, with a "Beowulf" excerpt detailing happier times for the title character, as he recounts vanquishing his adversaries with a big sword. The translation by J. Lesslie Hall is posted to the Project Gutenberg website:

"Beowulf spake, offspring of Ecgtheow:

'Lo! we blithely have brought thee, bairn of Healfdene,

Prince of the Scyldings, these presents from ocean

Which thine eye looketh on, for an emblem of glory.

I came off alive from this, narrowly 'scaping:

In war 'neath the water the work with great pains I

Performed, and the fight had been finished quite nearly,

Had God not defended me. I failed in the battle

Aught to accomplish, aided by Hrunting,

Though that weapon was worthy, but the Wielder of earth-folk

God was fighting with me.

Gave me willingly to see on the wall a

Heavy old hand-sword hanging in splendor

(He guided most often the lorn and the friendless),

That I swung as a weapon. The wards of the house then

I killed in the conflict (when occasion was given me).

Then the battle-sword burned, the brand that was lifted,

As the blood-current sprang, hottest of war-sweats;

Seizing the hilt, from my foes I offbore it;

I avenged as I ought to their acts of malignity,

The murder of Danemen. I then make thee this promise,

Heorot is freed from monsters.

Thou'lt be able in Heorot careless to slumber

With thy throng of heroes and the thanes of thy people

Every and each, of greater and lesser,

And thou needest not fear for them from the selfsame direction

As thou formerly fearedst, oh, folk-lord of Scyldings,

End-day for earlmen.' ”

John R. Ellement of the Globe Staff and Globe correspondent Terry Byrne contributed to this report. Travis Andersen can be reached at travis.andersen@globe.com. Follow him on Twitter [@TAGlobe](https://twitter.com/TAGlobe).

©2019 Boston Globe Media Partners, LLC